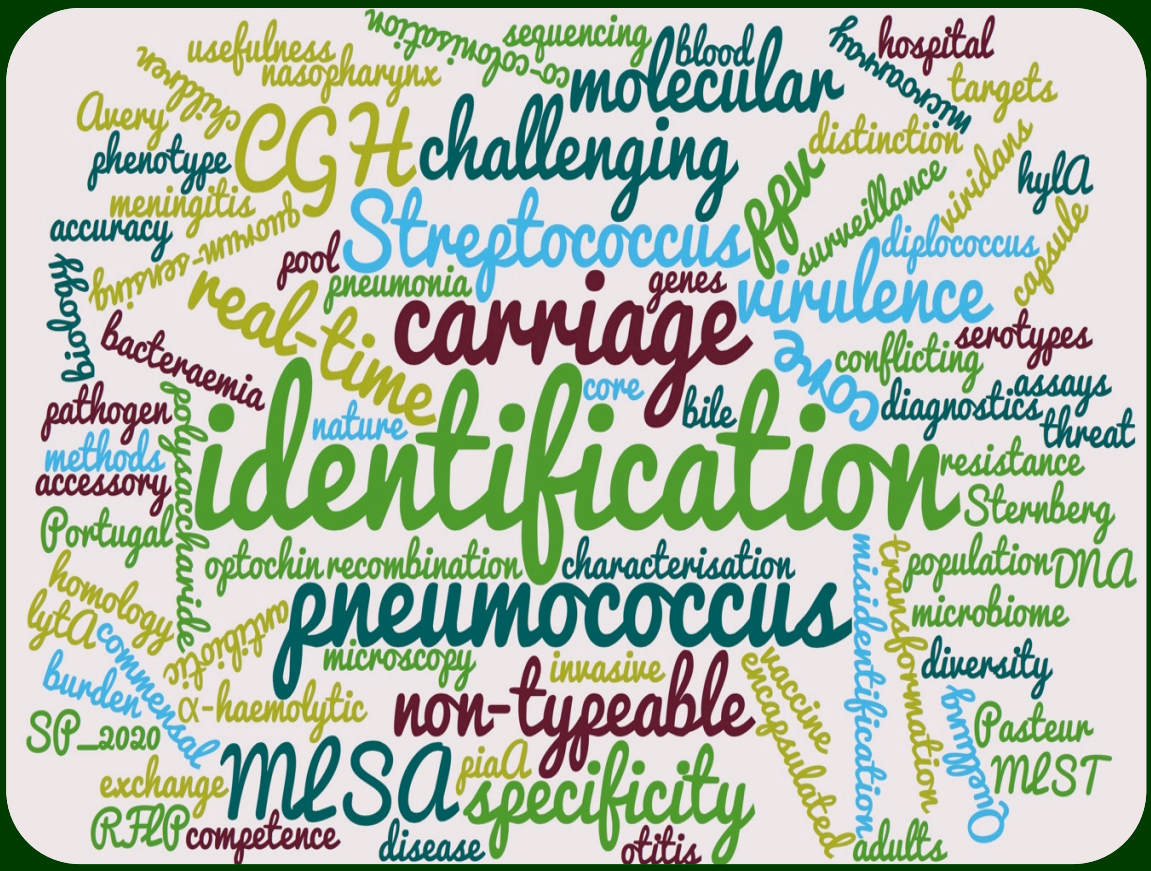
Studies on non-typeability and molecular identification of the pneumococcus

Débora A. Tavares



Dissertation presented to obtain a Ph.D degree in Biology | Molecular Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
March, 2016

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL
Knowledge Creation

Studies on non-typeability and molecular identification of the pneumococcus

Débora A. Tavares

Dissertation presented to obtain a Ph.D degree in Biology | Molecular Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras, March, 2016

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL

Knowledge Creation

Financial support: Fundação para a Ciência e a Tecnologia through grant SFRH/BD/70147/2010, awarded to Débora A. Tavares

Second Edition, May 2016

ISBN: 978-989-20-6369-0

Supervisors:

Raquel Sá-Leão

Hermínia de Lencastre

Examiners:

Birgitta Henriques-Normark

Markus Hilty

Mónica Serrano

Ana Madalena Ludovice

Dissertation presented on March 21, 2016, to obtain a Ph.D degree in Biology | Molecular Biology

by Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa.

Acknowledgements

To Raquel Sá-Leão, my supervisor. What can I say? Sometimes the most difficult person to acknowledge is the one that most credit deserves. My admiration for you as a person, a scientist, and a mentor is nothing new. Even so, I feel compelled to mention that it is amazing to feel your care and friendship, your willingness to see us exceed our expectations, your constant challenge, your enthusiastic curiosity, your inspiring spirit, and your love for science. Thank you for all the guidance and support, but most of all, thank you for putting an end to it.

To Professor Hermínia de Lencastre, my co-supervisor, for the opportunity of doing my Ph.D. at the Microbiology of Human Pathogens Unit, for her demand for excellence, and for the valuable advices through the years.

To all my former and current colleagues at the Microbiology of Human Pathogens Unit, especially the 'Vamos equipa' crew, for all the laughs and shared moments. To Carina Valente, Cristina Paulo, Sónia Almeida, Sofia Félix, and Sara Handem for the critical discussions of my work.

To our collaborators Peter Hermans, Hester Bootsma, Mark Eleveld, and Aldert Zomer, without whom part of this work would not have been performed. Thank you for the great experience that was to work at the Laboratory of Paediatric Infectious Diseases at the Radboud University Nijmegen Medical Centre, the Netherlands. To our collaborator Jason Hinds from the Bacterial Microarray Group at St George's, University of London, UK.

To Adriano Henriques and Isabel Couto, members of my Ph.D. thesis committee, for precious advice and discussion of my work.

To the “DCC team” and all the people participating in our colonisation surveillance studies.

To ITQB, for the excellent working conditions, and to Fundação para a Ciência e a Tecnologia, for financial support.

Last but not least, to my family and friends. Nothing in my life would ever be possible without you and life is only this amazing because I get to share it with you. Thank you for all your support and your understanding, but above all, thank you for your unconditional love.

Abstract

Pneumococcus is a major human pathogen. Its main virulence factor is the capsule, which is of polysaccharide nature. The detection of the capsule using specific antisera is used to identify pneumococcus. Also, differences in structural and antigenic properties of the polysaccharides composing the capsule have been used to classify pneumococcus into serotypes.

Pneumococcus that do not express a capsule are maintained in nature. In comparison with encapsulated pneumococci, non-encapsulated pneumococci (NT) have been poorly studied. Additionally, the distinction between NT and closely related *Streptococcus* species can be difficult, even when molecular methods are used.

The aims of this thesis were to gain insights into the genetic basis of non-typeability and genomic content and diversity of NT circulating in Portugal, and to improve molecular identification of pneumococcus by real-time PCR.

To address these goals three studies were performed: (i) the characterisation of the NT population circulating in Portugal, (ii) the characterisation of α-haemolytic isolates of the viridans group of *Streptococcus* of ambiguous identification by widely used pneumococcal identification methods, and (iii) the evaluation of the performance of currently in use and novel real-time PCR strategies to identify pneumococcus.

In the first study, 52 NT strains representing the lineages circulating in Portugal between 1997 and 2007 were characterised to gain insights into their genetic

content and mechanisms leading to non-typeability and genetic diversity. By sequencing of the capsular region, NT were found to be an homogeneous group belonging to *cps* type NCC2, as all strains harboured both *aliB*-like ORF1 and *aliB*-like ORF2 genes. By comparative genomic hybridisation with a microarray covering the genome of 10 pneumococcal strains, the core genome of NT was found to be essentially similar to that of encapsulated strains, with competence genes and most virulence genes being present. The intraclonal variation found among NT could not be entirely explained by the presence of mobile elements.

The second study originated from the identification, during previous pneumococcal carriage surveillance studies, of 11 α-haemolytic streptococcal isolates displaying conflicting or novel results by widely accepted pneumococcal identification methods. To understand the genetic basis for the unexpected results obtained for *lytA*-BsaAI-RFLP, the isolates were further characterised. The usefulness of the *lytA*-CDC real-time PCR, currently recommended by the WHO for the culture-independent identification of pneumococcus, was also assessed for these isolates. Use of the MLST scheme for pneumococcus and the MLSA scheme for the viridans group of *Streptococcus* led to the identification of four pneumococci among the 11 isolates. Three of the four pneumococci had a 60bp deletion in *lytA*, resulting in a new *lytA*-BsaAI-RFLP pattern but not in misidentification by the *lytA*-CDC real-time PCR. The fourth pneumococcal strain harboured a *lytA* homologue, resulting in misidentification by both *lytA*-BsaAI-RFLP and the *lytA*-CDC real-time PCR. The remaining seven isolates were identified as five *S. mitis* and two *S. pseudopneumoniae*. All *S. mitis* had point mutations in the *lytA* homologue, resulting

in two new *lytA*-BsaAI-RFLP patterns. These isolates were correctly identified by the *lytA*-CDC real-time PCR. The two *S. pseudopneumoniae* harboured (pneumococcal) *lytA*, resulting in misidentification by both *lytA*-BsaAI-RFLP and the *lytA*-CDC real-time PCR.

In the third study, a collection of close to 600 streptococcal isolates was tested to evaluate the performance of real-time PCR assays in the identification of pneumococcus. This collection included, as control strains, 150 pneumococci and 31 strains of *Streptococcus* species other than pneumococcus. Four real-time PCR assays were tested: the *lytA*-CDC real-time PCR assay and three other assays targeting *piaB*, *hylA*, and SP_2020. The *piaB* real-time PCR assay was previously used by others in parallel with the *lytA*-CDC to increase specificity of the assay. The assays targeting *hylA* and SP_2020 were designed under the scope of this thesis. The best real-time PCR assays for the identification of pneumococcus were SP_2020 and *lytA*-CDC, with a specificity of 100% and positive predictive value of 99.3%. and 98.7, respectively (p=0.564).

Overall, the work presented in this thesis provides evidence supporting the inclusion of non-typeable pneumococci in the species *S. pneumoniae*, alerts for the existence of streptococcal isolates for which an accurate identification can be challenging even by some of the widely accepted molecular methods currently in use, and contributed with a new assay for the identification of pneumococcus by real-time PCR.

Resumo

O pneumococo é um dos mais importantes patogéneos humanos. O seu principal factor de virulência é a cápsula, de natureza polissacárida, cuja detecção com anticorpos específicos é utilizada na sua identificação. As diferenças estruturais e antigénicas dos polissacáridos constituintes da cápsula permitem a divisão de pneumococos em serótipos.

Existem pneumococos que não possuem cápsula – designados pneumococos não capsulados – que estão pouco estudados. Devido à ausência de cápsule, a distinção entre pneumococos não capsulados e membros das espécies de *Streptococcus* filogeneticamente mais próximas de pneumococos pode ser difícil, mesmo quando feita através de métodos moleculares.

Esta tese teve como objectivos aprofundar o conhecimento relativo aos pneumococos não capsulados, nomeadamente as alterações genéticas que causam a ausência de expressão de cápsula, o conteúdo genético e a diversidade de pneumococos não capsulados em circulação em Portugal, assim como aperfeiçoar a identificação molecular de pneumococos por PCR em tempo real.

Para alcançar estes objectivos, foram efectuados três estudos: (i) a caracterização da população de pneumococos não capsulados em circulação em Portugal, (ii) a caracterização de isolados α-hemolíticos do grupo viridans do género *Streptococcus* com características ambíguas que tornam difícil a sua identificação e (iii) a avaliação de resultados de estratégias de identificação de pneumococos por PCR em tempo

real. Neste último ponto, foram avaliadas tanto estratégias actualmente utilizadas como novas estratégias desenvolvidas no âmbito desta tese.

No primeiro estudo, foram caracterizadas 52 estirpes de pneumococos não capsulados representativas das linhagens em circulação em Portugal entre 1997 e 2007, com o objectivo de conhecer melhor o conteúdo genético destas estirpes e compreender os mecanismos responsáveis pela ausência de cápsula pela diversidade genética encontrada. Através da sequenciação da região capsular foi possível perceber que os pneumococos não capsulados estudados representam um grupo homogéneo de estirpes pertencentes ao tipo *cps* NCC2, uma vez que todas continham os genes *aliB*-like ORF1 e *aliB*-like ORF2. Por estudos de hibridação genómica comparativa com um *microarray* que inclui o genoma de 10 estirpes de pneumococos, foi possível observar que o genoma essencial (de *core genome*) dos pneumococos não capsulados é, na sua essência, semelhante ao de estirpes capsuladas de pneumococos. Os genes de competência e a maioria dos genes de virulência presentes em estirpes capsuladas foram detectados nos pneumococos não capsulados. A variabilidade intra-clonal encontrada nas NT não pôde, no entanto, ser integralmente explicada pela presença de elementos móveis, com base na metodologia utilizada.

O segundo estudo teve origem na identificação, no decorrer de estudos de vigilância de colonização por pneumococos, de 11 isolados α-hemolíticos do género *Streptococcus* com resultados díspares ou novos quando foram usados diferentes métodos de identificação de pneumococos. Para procurar perceber os resultados inesperados obtidos por *lytA*-BsaAI-RFLP, foi feita uma caracterização mais

detalhada destes isolados. A pertinência de se utilizar o ensaio do CDC de PCR em tempo real para o gene *lytA* (*lytA*-CDC), actualmente recomendado pela OMS para identificação de pneumococos por métodos não dependentes de cultura, foi também testada para estes isolados. A utilização do esquema de MLST para pneumococos e do esquema de MLSA para *Streptococcus* do grupo viridans resultou na identificação de quatro pneumococos entre os 11 isolados. Verificou-se que três das quatro estirpes de pneumococos tinham uma deleção de 60bp no gene *lytA*, o que resultou num novo padrão de *lytA*-BsaAI-RFLP, mas não afectou a correcta identificação pelo *lytA*-CDC. A quarta estirpe de pneumococos possuía um homólogo do gene *lytA* (não possuindo a cópia "nativa de pneumococcus), o que levou a uma incorrecta identificação tanto por *lytA*-BsaAI-RFLP como pelo *lytA*-CDC. Os restantes sete isolados foram identificados como cinco *S. mitis* e dois *S. pseudopneumoniae*. Em todos os *S. mitis* foram detectadas mutações pontuais num gene homólogo de *lytA*, resultando em dois novos padrões de *lytA*-BsaAI-RFLP. Estes isolados foram correctamente identificados pelo *lytA*-CDC. Nos dois *S. pseudopneumoniae* foi detectado o gene *lytA*, o que resultou na incorrecta identificação tanto por *lytA*-BsaAI-RFLP como pelo *lytA*-CDC.

No terceiro estudo estudou-se uma colecção de cerca de 600 isolados do género *Streptococcus* para avaliar o desempenho de vários ensaios de PCR em tempo real na identificação de pneumococos. Esta colecção incluiu, como estirpes controlo, 150 pneumococos e 31 estirpes de espécies do género *Streptococcus* que não pneumococos. Foram feitos quatro ensaios de PCR em tempo real: o *lytA*-CDC e três outros tendo como alvo os genes *piaB*, *hylA* e SP_2020. O ensaio de PCR em tempo

real *piaB* é usado por outros grupos de investigação em paralelo com o *lytA*-CDC para aumentar a especificidade do ensaio. Os ensaios tendo como alvos os genes *hylA* e SP_2020 foram desenhados no âmbito desta tese. Os ensaio em que se obtiveram melhores resultados para a identificação de pneumococos foram o ensaio SP_2020 e o ensaio *lytA*-CDC, com uma especificidade de 100% e um valor preditivo positivo de 99.3%. e 98.7%, respectivamente (p=0.564).

No seu conjunto, o trabalho apresentado nesta tese apoia a inclusão de pneumococos não capsulados na espécie *S. pneumoniae*, alerta para a existência de isolados do género *Streptococcus* para os quais a correcta identificação da espécie pode ser difícil, mesmo utilizando métodos amplamente aceites de entre os actualmente disponíveis, e contribui com um novo ensaio para a identificação de pneumococos por PCR em tempo real.

Thesis outline

The purpose of the work presented in this thesis was to gain insights into the genetic basis of non-typeability and genomic content and diversity of non-typeable pneumococci circulating in Portugal, and to improve molecular identification of pneumococcus by real-time.

Chapter 1 is a general introduction presenting important aspects of the biology of the pneumococcus relevant for the scope of this thesis. In particular, epidemiology, the polysaccharide capsule, and identification methods are among the topics covered.

Chapter 2 describes the genetic characterisation of the non-typeable pneumococcal population in circulation in Portugal between 1997 and 2007. Genomic DNA of the samples was analysed with a microarray covering the genome of ten pneumococcal strains and the capsular region was sequenced.

Chapter 3 and **Chapter 4** describe studies addressing the molecular identification of pneumococcus.

Chapter 3 describes the characterisation of 11 isolates of the viridans group of *Streptococcus* presenting conflicting or novel results by widely accepted pneumococcal identification methods. The *lytA* gene was sequenced, the usefulness of using molecular methods targeting this gene was assessed, and multilocus sequence analysis was used to infer the species.

Chapter 4 describes a study conducted to evaluate currently available and novel real-time PCR assays for the identification of pneumococcus. Four real-time PCR

assays were tested with a collection of close to 600 samples, including 150 pneumococcus and 31 strains from other *Streptococcus* species.

Chapter 5 presents general conclusions of the studies conducted in this thesis and enumerates several questions that remain unanswered and could be the focus of future research.

Chapters 2, 3, and 4 can be read independently.

Chapter 2 is a reproduction of the following publication: **D. A. Tavares*, A. S. Simões*, H. J. Bootsma, P. W. M. Hermans, H. de Lencastre, and R. Sá-Leão.** 2014. Non-typeable pneumococci circulating in Portugal are of cps type NCC2 and have genomic features typical of encapsulated isolates. BMC Genomics 15: 863-77. *Equal contribution.

Chapter 3 is a reproductin of the following publication: **A. S. Simões*, D. A. Tavares*, D. Rolo, C. Ardanuy, H. Goossens, B. Henriques-Normark, J. Linares, H. de Lencastre, and R. Sá-Leão.** 2016. *lytA*-based identification methods can misidentify *Streptococcus pneumoniae*. Diagn. Microbiol. Infect. Dis. doi:10.1016/j.diagmicrobio. 2016.03.18 [Epub ahead of print]. *Equal contribution.

Chapter 4 is nearly ready for submission.

Table of contents

Chapter 1

Introduction

The pneumococcus

The sugar-coated bacteria in more than a century of pioneering research

Pneumococcus was isolated for the first time in 1880 by G. M. Sternberg after inoculation of his own saliva into mice. In the same year, pneumococcus was also independently recovered by L. Pasteur after inoculation of mice with saliva from a child who had died of rabies. Lancet-shaped pairs of coccoid bacteria were obtained and the presence of a capsule surrounding the diplococcal form of the organism was observed (reviewed in Watson *et al.*, 1993; Lopez, 2006) (Figure 1). Two years later, pneumococcus was identified as the major cause of bacterial pneumonia in humans and, soon after, it was also associated with meningitis, endocarditis, arthritis, and otitis media (reviewed in Austrian, 1999; Lopez, 2006). At the beginning of the twentieth century, the burden of pneumococcal pneumonia was so high that the disease was termed “The captain of the men of death” (reviewed in Austrian, 1999). Such a threat to human life soon attracted the attention of the scientific community to the study of the biology of pneumococcus and its ability to cause disease, in a demand to treat and control pneumococcal disease.

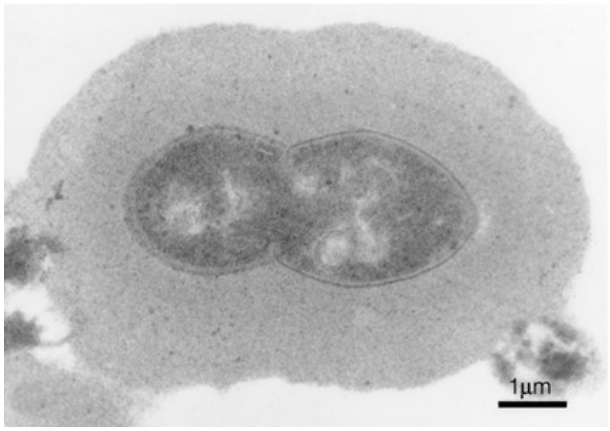


Figure 1. The pneumococcus (adapted from Kim *et al.*, 1999).

To facilitate the visualisation of pneumococcus in histological sections of the lungs, H. C. Gram developed in 1884 the Gram staining, the phenotypic method that is still in use today and that divides most bacteria in two groups: Gram-positive and Gram-negative (reviewed in Austrian, 1960). Also important was the identification of characteristics that differentiated pneumococcus from other bacteria by F. Neufeld. In 1900, Neufeld developed the bile solubility test and, in 1902, the Quellung reaction (reviewed in Lopez, 2006).

Pneumococcus also played an historical role in the development of immunology and design of vaccines. In 1891, G. Klemperer and F. Klemperer showed that serum from rabbits injected with heat-killed pneumococci or culture filtrates contained factors that conferred immunity to reinfection with the same strain but not necessarily with different ones. More importantly, serum from a previously immunised animal conferred protection against primarily pneumococcal infection (reviewed in Watson *et al.*, 1993). Two years later, B. Issaeff demonstrated that the serum did not have bactericidal properties but instead was promoting the uptake of pneumococcus by phagocytic cells (reviewed in Watson *et al.*, 1993). In 1913, serum treatment of pneumococcal pneumonia started to be used at the Rockefeller Institute (later the Rockefeller University) and case-fatality rates were reduced to c.a. 20% (reviewed in Austrian, 1999). In 1923, M. Heidelberg and O. T. Avery identified the soluble substances of pneumococcus as polysaccharides and showed that they had antigenic properties similar to the ones exclusively attributable, at that time, to proteins (Heidelberger and Avery, 1923; reviewed in Watson *et al.*, 1993).

Simultaneously to the beginning of serotherapy, J. Morganroth and M. Kaufmann reported, in 1911, the use of ethylhydrocupreine (optochin, a derivative of the antimalarial quinine) to treat experimentally infected mice, one of the first examples of the use of a specific antimicrobial agent as therapy to treat serious bacterial infection (reviewed in Watson *et al.*, 1993; Lopez, 2006). However, the use of optochin was soon abandoned due to its optic toxicity (reviewed in Watson *et al.*, 1993). Sulfapyridine was also the treatment of choice for a short period until the introduction of penicillin in the early 1940s. The use of penicillin reduced the overall fatality rate of pneumococcus to 5-8%, an improvement in comparison to the reduction to the vicinity of 20% previously accomplished with serotherapy (reviewed in Lopez, 2006). The impact of the introduction of antimicrobials was so impressive that the study of microorganisms was considered by some a waste of time. The battle against infectious diseases was declared won. But this was a sour victory, as antimicrobial resistance soon emerged and spread, renewing the interest in the study of serious infectious diseases (reviewed in Austrian, 1999).

Perhaps the first antimicrobial resistant organism recovered from an animal was an optochin resistant pneumococcus isolated from treated mice in 1912. Pneumococci expressing significantly increased resistance to optochin were also recovered from the blood of treated patients in 1917 and 1918 (reviewed in Austrian, 1999). In 1943, the first sulphonamide-resistant isolates were reported by Tillett, *et al*. More importantly, it was also in 1943 that L. H. Schmidt and C. L. Sessler described the isolation of penicillin resistant pneumococci from mice treated for experimental pneumococcal infection. Two years later, penicillin resistant pneumococci were

recovered by K. R. Eriksen after *in vitro* exposure to penicillin (reviewed in Austrian, 1999). Because these observations had little impact on clinical practice, antimicrobial resistance was mostly overlooked until the 1980s, when infections caused by antimicrobial-resistant pneumococci started to become an increasing concern and products encoded by bacterial genes and plasmids were found to interfere with the available antimicrobials (reviewed in Austrian, 1999; Lopez, 2006).

Attention was then again turned to the development of efficient vaccines to prevent pneumococcal disease. Ironically, studies for the development of pneumococcal vaccines had been initiated in 1911 (one year before the isolation of the first antimicrobial resistant pneumococcus) with the work of A. E. Wright and F. S. Lister suggesting that inoculation of dead pneumococci might elicit protection against pneumococcal infection in humans. At the time, Lister also reasoned that immunisation of half the members of a closed population would benefit the unimmunised half by preventing the spread of pneumococcus in the population (reviewed in Watson *et al.*, 1993; Austrian, 1999). In 1926, capsular polysaccharides were isolated and a number of studies were conducted in the 1930s and 1940s on the effectiveness of vaccines aimed to prevent pneumococcal disease, all showing that healthy adult volunteers were protected against pneumococcal infection by vaccines that stimulated the immune system to produce antibodies against pneumococcus (reviewed in Watson *et al.*, 1993). In fact, in 1937, active vaccination with a relevant subcellular bacterial fraction was used for the first time in a program of mass vaccination to abort an outbreak of pneumonia in a state hospital. Ten years later, pneumococcal vaccines containing two and later three type-specific

polysaccharides were demonstrated efficacious in an elderly cohort, leading to the commercial production of two hexavalent polysaccharide vaccines. However, these vaccines were soon withdrawn from the market as they became available when all interest was turned to the effectiveness of antimicrobials (reviewed in Austrian, 1999). Close to twenty years then passed until studies by R. Austrian, in the 1960s, renewed the interested in pneumococcal polysaccharide vaccines. Austrian showed that, unlike what was stated at the time, treatment with antimicrobials did not result in a significant decline in the incidence of pneumococcal disease. In fact, case fatality rates were shown to reach 17% for adults with uncomplicated bacteraemic pneumococcal pneumonia treated with penicillin, exceeding 25% in high-risk populations (Austrian and Gold, 1964). Studies by Austrian culminated with the introduction of a 14-valent vaccine in 1977, containing the polysaccharide components of each of the 14 most common pneumococcal serotypes, responsible for 80% of cases of pneumococcal disease. In 1983, this vaccine was expanded to the 23-valent composition still available today (PPV23), the most complex vaccine ever administered to humans (reviewed in Watson *et al.*, 1993).

But the finding with the highest impact on biology to arise from the study of pneumococcus (and bacteriology) was undeniably that DNA is the genetic material (Avery *et al.*, 1944). Studies leading to this historical finding started in 1928 when F. Griffith reported that co-infection of mice with live avirulent bacteria expressing a rough phenotype and dead virulent strains expressing a smooth phenotype resulted in the death of mice by pneumococci that had been transformed to the smooth phenotype. In this first report of natural genetic transformation, Griffith termed the

substance responsible for the genetic exchange in bacteria resulting in an unexpected phenotypic change “the transforming principle” (reviewed in Watson *et al.*, 1993; Lopez, 2006). In Avery’s laboratory, it had been recently demonstrated that the differences between pneumococcal strains resulted exclusively from differences in the structure of the polysaccharides in their capsules and not from more complex biological factors. So the idea that strains were not fixed entities, as regarded by many at the time, was received in Avery’s laboratory with apprehension (reviewed in Russell, 1988). But carbohydrates could not be the substance from the dead strain incorporated to modify the live one because carbohydrates were thermostable and transformation would not occur if the dead strain was subject to temperatures higher than 80°C (reviewed in Russell, 1988). Nevertheless, Griffiths’ studies were only recreated and extended by Avery and his team (C. M. Mac-Leod and M. McCarty) after others had confirmed the occurrence of transformation (reviewed in Russell, 1988). In the experiments performed in Avery’s laboratory, previously unencapsulated (rough) strains would acquire a capsule through the incorporation of some component from crude extracts of the smooth variant becoming, in McCarty’s words, a “sugar-coated bacteria”. In 1944, Avery, Mac-Leod, and McCarty determined that Griffith’s “transforming principle” was DNA. In the words of Avery, “it means that nucleic acids are not merely structurally important but functionally active substances in determining the biochemical activities and specific characteristics of cells and that by means of a known chemical substance it is possible to produce predictable and hereditary changes in cells. This is something that has long been the dream of geneticists.” (reviewed in Austrian, 1999). This finding marked the origin of molecular biology. Two decades later, while studying

pneumococcal transformation A. Tomasz described for the first time the fascinating phenomenon of quorum sensing, that results in communication between bacteria (Tomasz, 1965).

Pneumococcus and humans, for better or worse

Although pneumococcus has always been the subject of great interest due to its undeniable clinical importance, pneumococcus is mainly a commensal of the human upper respiratory tract (Bogaert *et al.*, 2004). Nasopharyngeal colonisation is very frequent in young children (with percentages of carriage reaching 60%) becoming less frequent at older ages (Aniansson *et al.*, 1992; Hussain *et al.*, 2005; Nunes *et al.*, 2005). In fact, the nasopharynx of young children has been considered the main reservoir of pneumococcus (Sá-Leão *et al.*, 2000; Bogaert *et al.*, 2004), while carriage in adults and the elderly has been regarded as negligible (Regev-Yochay *et al.*, 2004; Almeida *et al.*, 2014).

The duration of pneumococcal colonisation usually lasts a couple of weeks, but colonisation periods can reach several months, especially in younger children (Sá-Leão *et al.*, 2008; Abdullahi *et al.*, 2012). Also, children can acquire several different pneumococcal strains over time, with co-occurrence of several strains not being unusual (Valente *et al.*, 2012), and less immunogenic serotypes tend to be carried for longer periods of time (reviewed in Donkor, 2013). Because pneumococcus is mainly air-borne transmitted from person to person, a new colonisation event greatly depends on close contact with a pneumococcal carrier (Bogaert *et al.*, 2004). For this

reason, crowed places such as day care centres, prisons, and military camps increase the chance of pneumococcal transmission and, consequently, of pneumococcal colonisation (Sá-Leão *et al.*, 2008; reviewed in Donkor, 2013). Transmission of pneumococcus has also been associated with sharing of bottles by young adults (Levine *et al.*, 2012). Also, although transmission through environmental surfaces has not been described, pneumococcus has been observed to survive desiccation for weeks (Walsh and Camilli, 2011).

Pneumococcal colonisation is also a pre-requisite for disease so pneumococcal carriers act not only as the reservoir but also as the source of disease-causing strains (Simell *et al.*, 2012; reviewed Bogaert *et al.*, 2004). However, although almost every child will have experienced at least one episode of colonisation in young age, only a small proportion of carriers goes through an evolution from colonisation to disease (reviewed in Donkor, 2013). Risk factors for pneumococcal disease include extreme of ages (below two years of age or above 65) and underlying medical conditions such as chronic diseases and immunocompromising conditions (reviewed in Bogaert *et al.*, 2004). Nevertheless, pneumococcus is still a major public health problem, with more than 11 million episodes of serious pneumococcal disease estimated to occur annually in the world (O'Brien *et al.*, 2009). As a consequence, more than 800,000 children under the age of five die every year of pneumococcal disease (c.a. 10% of all deaths among children in this age group), especially in African and Asian countries (O'Brien *et al.*, 2009).

In the US, pneumococcal disease causes 4 million disease episodes and 22,000 deaths annually, with more than 1 million infections and 7,000 deaths being caused

by antimicrobial-resistant pneumococcus. Among children, otitis media is the most common pneumococcal disease, causing 1.5 million infections. Another important pneumococcal disease is pneumonia, with nearly 160,000 children younger than five years of age and over 600,000 adults seeing a doctor or being admitted to the hospital. This disease accounts for c.a. 70% of all medical cost for treatment of pneumococcal disease. In 30% of severe cases, pneumococcus is fully resistant to one or more clinically relevant antimicrobials, such as penicillin and erythromycin, resulting in c.a. 30,000 additional doctor visits, nearly 20,000 hospitalisations, and c.a. 100 million dollars in excess costs. The majority of invasive disease and deaths due to pneumococcal disease occurs in adults over 65 years of age (CDC, 2013).

The development of pneumococcal vaccines has been a major improvement in the combat of pneumococcal infection through its prevention. Although 97 pneumococcal serotypes have been identified, only a few are responsible for most cases of disease (Henrichsen, 1995; Ko *et al.*, 2013). When PPV23 was designed, this was already taken into account as this vaccine included the serotypes most commonly causing disease at the time (Robbins *et al.*, 1983). One of the drawbacks of this vaccine is that children younger than two years of age, a population at increased risk of developing pneumococcal disease, shows a poor immunogenic response to polysaccharides (Leinonen *et al.*, 1986; O'Brien *et al.*, 1996).

In 2000, a paediatric pneumococcal conjugate vaccine (PCV7) was introduced in the US national immunisation program to be administered to children younger than two years of age. This vaccine allied capsule polysaccharides of the seven most prevalent serotypes causing invasive pneumococcal disease among young children in the US to

CRM197, a non-toxic variant of the diphtheria toxin (CDC, 2000; Hausdorff *et al.*, 2000a; Hausdorff *et al.*, 2000b). In Portugal, PCV7 was introduced in 2001 and although it was not included in the National Immunisation Plan nor was it reimbursed by the state, close to 80% of the target population was being vaccinated by 2007 [data from IMS and INE (National Statistics Institute)]. Following the introduction of PCV7 in the US, there was a significant decrease in the incidence of invasive pneumococcal disease caused by the serotypes included in the vaccine in both vaccinated children and unvaccinated people of all ages (herd immunity), even though an increase in the numbers of cases caused by serotypes not included in the vaccine was reported (serotype replacement) (Whitney *et al.*, 2003; Hicks *et al.*, 2007; Kellner, 2011). A similar serotype replacement phenomenon was observed in colonisation studies, although rates of carriage remained unchanged (Sá-Leão *et al.*, 2009).

Although serotype formulation of PCV7 was chosen based on prevalence of invasive pneumococcal disease in children, the serotypes included in PCV7 were also associated with the majority of antimicrobial resistance in pneumococcus (Klugman, 1990; Dagan and Klugman, 2008). Also, vaccination with PCV7 led to a reduction in antimicrobial use (reviewed in Dagan and Klugman, 2008). Thus, with the simultaneous reduction of the serotypes carrying antimicrobial resistance and antimicrobial use following the introduction of PCV7, a decrease in antimicrobial resistance was expected. However, reports differ, with only some regions reporting a decrease in antimicrobial resistance (reviewed in Dagan and Klugman, 2008; Song *et al.*, 2012; Henriques-Normark and Tuomanen, 2013). The expected decrease in

antimicrobial resistance was not consistently observed in all countries where PCV7 was introduced due to an increase in antimicrobial resistance among serotypes not included in PCV7 observed both in cases of disease and in colonisation studies (Richter *et al.*, 2009; Sá-Leão *et al.*, 2009; Simões *et al.*, 2011a; Song *et al.*, 2012).

Currently, one polysaccharide pneumococcal vaccine (PPV23) and two pneumococcal conjugate vaccines (PCV10 and PCV13 in replacement of PCV7 and recently approved for all ages) are available. In Portugal, PCV13 was included in the National Immunisation Plan in June 2015. Although these vaccines have proven very effective in decreasing the burden of invasive pneumococcal disease in vaccinated populations, a universal pneumococcal vaccine, conferring immunity regardless of serotype and, consequently, not vulnerable to phenomena such as serotype replacement, has for long been an (so far) unreachable vision (reviewed in Simell *et al.*, 2012; Feldman and Anderson, 2014).

The polysaccharide capsule and life without it

The pneumococcal polysaccharide capsule is an external layer of saccharide repeat units that surrounds the cells, with each serotype having its own unique structure (Bentley *et al.*, 2006). In all but two serotypes (the exceptions being serotypes 3 and 37), capsular polysaccharide is covalently attached to the cell wall peptidoglycan (Sorensen *et al.*, 1990). The main function of the capsule during invasion is to shield pneumococcus against opsonisation and phagocytosis (Hyams *et al.*, 2010). However, the capsule has also been shown to increase invasiveness by influencing biofilm formation, sensitivity to neutrophil extracellular traps, and interaction with

the epithelium (reviewed in Geno *et al.*, 2015). Furthermore, the capsule is also important during colonisation, especially to repel mucus, delaying clearance (Nelson *et al.*, 2007).

Regardless of its importance, structural and immunogenic differences in the capsule result in different abilities of different serotypes to colonise and invade (invasive disease potential). These differences in invasive disease potential between serotypes suggest that host-pathogen interactions are serotype-dependent (reviewed in Geno *et al.*, 2015). In fact, some serotypes were found to be carried more than others and serotypes commonly associated with virulence among children are usually those with poorer immunogenicity at this age (reviewed in Geno *et al.*, 2015). These differences have been proposed to be related with different metabolic burden of capsular production and nutritional requirements of serotypes (Hathaway *et al.*, 2012).

The “capsular genome” is located between the genes *dexB* and *aliA* (not involved in the synthesis of capsular polysaccharide) and is referred to as the capsular polysaccharide synthesis (*cps*) locus (Garcia *et al.*, 2000). The first evidence for a same specific location of the *cps* locus, independently of the serotype being expressed, was published in 1959 by Austrian, *et al*. (Austrian *et al.*, 1959). Sequencing of *cps* loci of 90 serotypes also revealed a conserved locus structure beginning with *cpsA-D* widely conserved genes encoding for proteins involved in regulation of the capsule (Bentley *et al.*, 2006). Furthermore, capsular genes are predicted to be translated as a single operon (Guidolin *et al.*, 1994). This region can vary in size from 10kb to 30kb, with an average of 20kb (Bentley *et al.*, 2006). The

only exception seems to be serotype 37, as *tts*, encoding for a glucosyltransferase, is the only gene required for the production of capsular polysaccharide of serotype 37 and is located outside the *dexB-aliA* region (Llull *et al.*, 1999).

Besides the high diversity of serotypes recognised today, there are isolates that cannot be assigned to a serotype. These isolates, generally referred to as non-typeable, can fail to be serotyped because i) they are nonencapsulated; ii) they belong to a serotype still to be recognised; or iii) they have been misidentified as pneumococci. In this thesis, the term non-typeable (NT) will be used to refer exclusively to nonencapsulated pneumococci.

NT have been disregarded for many years, probably because of the small rough colonies formed by these strains when grown on agar plates. This morphology not only differs from the mucoid colonies characteristic of encapsulated pneumococci, but also resembles that of closely related *Streptococcus* species. Although generally considered less virulent than encapsulated strains, NT are frequent colonisers of the nasopharynx and have been associated with outbreaks of conjunctivitis (Martin *et al.*, 2003; Buck *et al.*, 2006; Sá-Leão *et al.*, 2006). Furthermore, increasing rates of NT nasopharyngeal colonisation after introduction of PCVs have been reported (Sá-Leão *et al.*, 2009).

Genetic studies of the *cps* locus have allowed the division of NT into two groups (Hathaway *et al.*, 2004; Park *et al.*, 2012; Salter *et al.*, 2012). Group I NT comprises pneumococcal isolates that although encoding for capsular genes do not express a capsule due to a defective capsular locus. These are usually related to encapsulated lineages based on multilocus sequence analysis (Scott *et al.*, 2012). The rare cases of

invasive disease traced back to NT seem to be mainly caused by Group I NT (Scott *et al.*, 2012; Park *et al.*, 2014). On the other hand, Group II NT comprises pneumococcal isolates encoding for AliB-like ORF 1 and AliB-like ORF2 or PspK, novel surface proteins suggested to facilitate adhesion to epithelial cells and colonisation (Hathaway *et al.*, 2004; Keller *et al.*, 2013). Group II NT isolates are frequent colonisers of the nasopharynx and the ones most frequently associated with conjunctivitis outbreaks (Martin *et al.*, 2003; Buck *et al.*, 2006; Hanage *et al.*, 2006; Sá-Leão *et al.*, 2006).

Recently, whole genome sequencing (WGS) of NT isolates revealed further detail on the genetics and epidemiological role of these isolates. Characterisation of a global collection of NT showed evidence of a long-standing classic lineage and multiple sporadic lineages. Isolates belonging to the classical lineage belonged mainly to ST344 and ST448, the predominant multilocus sequence types among NT, and had larger accessory genomes and higher rates of antimicrobial resistance than isolates belonging to the sporadic lineages. On the other hand, isolates belonging to the sporadic lineages clustered with encapsulated pneumococci (Hilty *et al.*, 2014). Furthermore, WGS of more than 3,000 carriage isolates from a 2.4km^2 refugee camp identified NT lineages as the ones with the highest frequencies of receipt and donation of recombined DNA fragments, suggesting a major role for these isolates in genetic exchange and adaptation of the species (Chewapreecha *et al.*, 2014).

The identification of pneumococcus

The relevance of the identification of pneumococcus

The correct identification of pneumococcus is important both for clinical practice and surveillance studies. The presumptive diagnosis of pneumococcal infections generally relies on the observation of Gram-positive encapsulated lanceolated cocci arranged in pairs or small chains in blood or other normally sterile body sites, with a definitive diagnosis obtained after isolation and characterisation of pneumococcus (CDC, 2013). However, a diagnosis is still not possible in most cases of invasive pneumococcal disease, especially in cases of pneumococcal pneumonia, as the interpretation of Gram stained sputum specimens can be challenged by the presence of normal nasopharyngeal flora and the isolation of pneumococcal cells may be impaired after antimicrobial therapy has started (Werno and Murdoch, 2008; CDC, 2013). This results in only 3-8% of positive blood cultures of hospitalised adults with pneumonia, in contrast to the more than 50% accomplished for meningitis (reviewed in Werno and Murdoch, 2008). Because most deaths caused by acute respiratory infections (the leading infectious cause of death in the world in both children and adults) are a consequence of pneumococcal pneumonia, empirical treatment is directed against pneumococcus when a diagnosis is not possible (Klugman *et al.*, 2008). This increases the selective pressure towards antimicrobial resistance, a worldwide concern given the emergence and dissemination of (especially) penicillin resistant pneumococci (CDC, 2013). Furthermore, the limitations in the identification of pneumococcus lead to difficulties in obtaining accurate data of invasive

pneumococcal disease burden and in evaluating the effectiveness of vaccination against pneumococcal disease (Werno and Murdoch, 2008).

In pneumococcal carriage surveillance studies, the correct identification of pneumococcus is important to estimate direct and indirect effects of vaccination on the pneumococcal population and to monitor antimicrobial resistance. The development and introduction of PCVs targeting an increasing number of pneumococcal serotypes is causing major changes in the pneumococcal population (Sá-Leão *et al.*, 2009). Also, misidentification of pneumococcal closely related species as pneumococcus can result in falsely increased resistance rates (Wester *et al.*, 2002; Richter *et al.*, 2008; Simões *et al.*, 2010). A fast cost-effective method that would allow the correct identification, serotyping, and determination of antimicrobial resistance profile of pneumococcus in just a few hours would then allow a more efficient antimicrobial prescription in the clinical practice, a more accurate estimation of invasive pneumococcal disease burden and vaccine efficacy, and an improved monitoring of antimicrobial resistance in the pneumococcal population.

Closely related streptococcal species and challenges in the identification of pneumococcus

Streptococcus is a large and diverse genus. The viridans group of *Streptococcus* includes non-pyogenic species that are mostly α-haemolytic and commensals of the oral cavity and pharynx, although it also includes some species that are commensals of the gastrointestinal and urogenital tract. Based on sequence homology of the 16S

rRNA, species of the viridans group have been further divided in four groups, including the mitis group to which pneumococcus belongs. Within the mitis group, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* are the closest relatives of pneumococcus (Hardie and Whiley, 1997; Facklam, 2002; Nakajima *et al.*, 2013) (Figure 2).

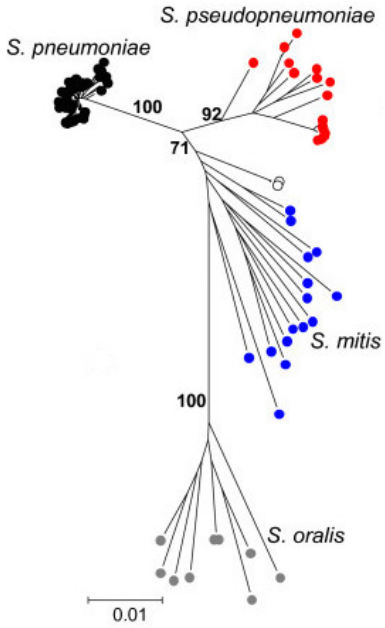


Figure 2. Phylogenetic relationship of pneumococcus and closely related species determined by multilocus sequence analysis (adapted from Bishop *et al.*, 2009).

Pneumococcus is considered the most important pathogen within the mitis group. However, S*. pseudopneumoniae*, *S. mitis*, and *S. oralis* have also been infrequently identified as cause of disease, especially in immunocompromised patients. *S. pseudopneumoniae* was described in 2004 and its clinical significance is less studied that that of *S. mitis* and *S. oralis* (Arbique *et al.*, 2004). Nevertheless, *S. pseudopneumoniae* has been recovered from both invasive and non-invasive infections and has been associated to infections of patients with chronic obstructive pulmonary disease (Harf-Monteil *et al.*, 2006; Keith *et al.*, 2006; Rolo *et al.*, 2013). *S.*

mitis and *S. oralis* have been associated with infection in neutropenic cancer patients and immediately after transplant surgery, sub-acute endocarditis, especially in patients with prosthetic valves, and septicaemia (Douglas *et al.*, 1993; Beighton *et al.*, 1994; Lucas *et al.*, 1997; Tunkel and Sepkowitz, 2002; Shelburne *et al.*, 2014).

Bacteria of the mitis group are competent for natural genetic transformation, which confers bacteria the capacity to actively uptake and heritably integrate extracellular DNA (Johnsborg *et al.*, 2007). During its commensal lifestyle, pneumococcus can frequently be found together with its close relatives. While sharing a niche, an opportunity is created for horizontal gene transfer to occur between these species. A good evidence of horizontal gene transfer between these species is the mosaic structure of gene sequences encoding for penicillin-binding proteins (Hakenbeck *et al.*, 1999). Further evidence is presented by the sharing of virulence genes among pneumococcus, *S. mitis*, and *S. oralis* (Whatmore *et al.*, 2000; Whalan *et al.*, 2006; Kilian *et al.*, 2008; Donati *et al.*, 2010; Johnston *et al.*, 2010; Morales *et al.*, 2015). As a consequence of the genetic exchange between pneumococcus and close relatives, species barriers can be blurred and a clear differentiation between these species may be challenging by both phenotypic and molecular methods.

Methods for the identification of pneumococcus

Traditional methods for the identification of pneumococcus rely on phenotypic characteristics of the species. Pneumococcus is a Gram-positive lancet-shaped coccus often occurring in pairs, with a capsule. When grown in blood-supplemented

media, pneumococcus forms round mucoid catalase negative α-haemolytic colonies. It is usually identified by its susceptibility to optochin and bile solubility (cell lysis in the presence of bile salts). Nonetheless, the definitive phenotypic identification method for the identification of pneumococcus is serotyping by the Quellung reaction, i.e., the serologic assignment of a capsular type or (serotype) based on immunogenic differences between pneumococcal capsules (Murray *et al.*, 2005).

Unlike pneumococcus, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* are resistant to optochin, bile insoluble, and negative for the Quellung reaction (Facklam, 2002; Arbique *et al.*, 2004; Ikryannikova *et al.*, 2011). However, non-pneumococcal strains expressing phenotypic traits characteristic of pneumococcus have been reported, providing evidence that misidentification of such strains as pneumococcus is possible by phenotypic identification (Fenoll *et al.*, 1994; Borek *et al.*, 1997; Arbique *et al.*, 2004; Simões *et al.*, 2010; Ikryannikova *et al.*, 2011; Rolo *et al.*, 2013). Also, some unusual pneumococci that are resistant to optochin or insoluble in bile have been isolated (Phillips *et al.*, 1988; Munoz *et al.*, 1990; Mundy *et al.*, 1998; Whatmore *et al.*, 2000; Pikis *et al.*, 2001; Obregon *et al.*, 2002; Nunes *et al.*, 2008). Furthermore, non-typeable pneumococci are negative for the Quellung reaction (Marsh *et al.*, 2010; Simões *et al.*, 2011b).

Although the phenotypic identification of pneumococcus by the combination of methods described above is routinely used successfully in hospitals and research laboratories, when ambiguous or unexpected results are obtained, confirmation by a genotypic method may be useful. Identification of bacteria by molecular methods has relied on the identification of ubiquitous genes that are not present in closely

related species or have significant sequence dissimilarity to allow the differentiation of species. A very useful bacterial molecular identification method is 16S rRNA sequencing. The rRNA nucleotide sequence is unique in that it harbours both regions that are very conserved among bacteria and regions that are hypervariable, allowing the design of both universal and species-specific primers (van Kuppeveld *et al.*, 1992; Greisen *et al.*, 1994). In recent years, this technique has also proven itself very useful in microbiome studies (Suau *et al.*, 1999; Sakamoto *et al.*, 2004; Bogaert *et al.*, 2011). However, within the viridans group of *Streptococcus*, which includes pneumococcus, 16S rRNA sequences exhibit 99% nucleotide homology, hampering the distinction between species (Kawamura *et al.*, 1995). Still, two assays aiming to distinguish pneumococcus from closely related streptococcal species based on specific 16S rRNA sequence differences have been recently proposed (El Aila *et al.*, 2010; Scholz *et al.*, 2012).

To overcome this problem, several pneumococcal genes, mostly virulence determinants, have been suggested as possible targets in pneumococcal identification strategies. Among them, *lytA* (encoding for the major pneumococcal autolysin), *ply* (encoding for pneumolysin), *psaA* (encoding for the pneumococcal surface adhesin A), and *pia* (encoding for the major iron ABC transport system of pneumococcus) have been investigated. A DNA probe for the identification of pneumococcus, targeting *lytA*, has been described more than 20 years ago and this gene is still used for the identification of pneumococcus (Pozzi *et al.*, 1989). Although homologs of *lytA* have been described in closely related species, nucleotide differences have allowed the design of assays for the identification of pneumococcus

targeting this gene, such as the method proposed by Llull *et al*. based on the characteristic BsaAI-RFLP signature of *lytA* that is useful to distinguish pneumococcus from closely related streptococcal species (Romero *et al.*, 2004; Llull *et al.*, 2006; Simões *et al.*, 2011b; Rolo *et al.*, 2013). Homologues of *ply* and *psaA* have also been described in pneumococcus close relatives but the specificity of PCR strategies to identify pneumococcus based on the amplification of these genes was lower than that of *lytA* (50% and 99% for *ply* and *psaA*, respectively), resulting in the misidentification of pneumococcus (Whatmore *et al.*, 2000; Jado *et al.*, 2001; Messmer *et al.*, 2004; Neeleman *et al.*, 2004). The major pneumococcal iron ABC transport system *pia* was first described as ubiquitous and specific to pneumococcus (Brown *et al.*, 2001). However, a more thorough study on the distribution of *piaA* among pneumococcus and closely related species showed evidence that although specific to pneumococcus, this transport system was not ubiquitous as it was not found in some non-typeable pneumococci (Whalan *et al.*, 2006).

Due to the genetic exchange between pneumococcus and closely related streptococcal species, a more reliable identification of pneumococcus may need to target more than one gene. Multilocus sequence typing (MLST) relies on nucleotide differences between internal fragments of seven housekeeping genes to assign an allelic profile or sequence type (ST). Each individual sequence is compared to all known alleles deposited at the MLST online database (www.mlst.net) for allelic number assignment. The seven allelic numbers form the allelic profile. This method has the advantages of being reproducible, comparable between laboratories, and to be an online database based tool. However, the costs associated with MLST make it

unsuitable for routine practice in many laboratories (Enright and Spratt, 1998; Maiden *et al.*, 1998). An MLST scheme has been successfully applied for the study of pneumococcus and is currently the gold standard for genotyping of pneumococcus (Enright and Spratt, 1998; Hanage *et al.*, 2005).

Furthermore, the sequences of multiple housekeeping genes from strains of closely related species have be concatenated and compared to generate clustering patterns to differentiate closely related species such as *Burkholderia* and *Neisseria* species (Hanage *et al.*, 2006). For the differentiation between pneumococcus and closely related species, two multilocus sequence analysis (MLSA) schemes have been used: the one relying on the MLST scheme for pneumococcus (www.mlst.net) and a second one targeting seven housekeeping genes for species of the viridans group of *Streptococcus* (Hanage *et al.*, 2006; Bishop *et al.*, 2009; Simões *et al.*, 2010; Rolo *et al.*, 2013). Whole genome sequencing (WGS) is being increasingly used and an approach combining automated extraction of WGS information with MLST-extended schemes may improve the resolution between closely related species such as the ones composing the viridans group of Streptococcus (Sabat *et al.*, 2013).

A cost-effective method that has recently become widely accepted for routine bacterial identification in clinical diagnostics due to its broad species coverage and fast turn-around time is matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS). For MALDI-TOF MS bacterial identification, colonies formed in solid medium are detached and directly analysed. Bacterial identification can also be performed to organisms grown in blood culture, but in this case a separation step is required (reviewed in Drancourt, 2010). Species

identification is then performed by measuring whole-spectrum similarities between whole-cell mass spectra of bacterial cell or cell extracts and reference spectra of well-characterised strains (reviewed in Fenselau and Demirev, 2001). Commercially available systems provide a reliable identification for the majority of clinically relevant species (van Belkum *et al.*, 2012). However, the accurate differentiation between pneumococcus and closely related species is still challenging and several studies have been performed trying to overcome this issue (Werno *et al.*, 2012; Ikryannikova *et al.*, 2013; Chen *et al.*, 2015).

Another method that has gained popularity in recent years is real-time PCR. Once an ideal target is identified, real-time PCR has the potential to unite the desired qualities of reliability, reproducibility, cost-effectiveness, and speed. It can also be used independently of culture, allowing the identification of pneumococcus even when isolation of pneumococcus is not possible. Furthermore, as amplification and detection occur within a closed system, real-time PCR also decreases the risk of contamination.

Currently, the culture-independent method recommended by the WHO for the identification of pneumococcus is the CDC real-time PCR strategy targeting *lytA* (Carvalho Mda *et al.*, 2007; Satzke *et al.*, 2013). Other targets, such as *ply* (Kaijalainen *et al.*, 2005; Carvalho Mda *et al.*, 2007; Abdeldaim *et al.*, 2010), *psaA* (Carvalho Mda *et al.*, 2007) or Spn9802 (Abdeldaim *et al.*, 2008) have been tested for the identification of pneumococcus by real-time PCR, but performed poorer than *lytA* (Carvalho Mda *et al.*, 2007). However, the *lytA*-CDC real-time PCR assay is not 100% accurate and a second real-time PCR assay, targeting *piaB*, has been used in

parallel with that targeting *lytA* to increase the specificity of pneumococcal identification (Trzcinski *et al.*, 2013; Wyllie *et al.*, 2014).

Improving molecular identification of pneumococcus

Since there is no perfect method for the identification of pneumococcus, proposed strategies should be extensively tested before being implemented as routine procedures. Also, new target genes for the identification of pneumococcus should be identified and tested.

Aim of the work

The aim of the work presented in this thesis was to: (i) gain insights into the genetic basis of pneumococcal non-typeability and the genomic content and diversity of NT through the characterisation of a carriage collection of NT circulating in Portugal in a period of 11 years; (ii) determine the genetic basis for the unexpected results given by *lytA*-BsaAI-RFLP for 11 presumptive pneumococcal isolates and investigate the accuracy of the *lytA*-CDC real-time PCR assay in the classification of these isolates; and (iii) evaluate the performance of two previously described (*lytA*-CDC and *piaB*) and two novel (*hylA* and SP_2020) real-time PCR assays for the identification of pneumococcus.

References

Abdeldaim, G., B. Herrmann, P. Molling, H. Holmberg, J. Blomberg, P. Olcen and K. Stralin. (2010). Usefulness of real-time PCR for *lytA*, *ply*, and Spn9802 on plasma samples for the diagnosis of pneumococcal pneumonia. *Clin Microbiol Infect* **16**, 1135-41.

Abdeldaim, G.M., K. Stralin, P. Olcen, J. Blomberg and B. Herrmann. (2008). Toward a quantitative DNA-based definition of pneumococcal pneumonia: a comparison of *Streptococcus pneumoniae* target genes, with special reference to the Spn9802 fragment. *Diagn Microbiol Infect Dis* **60**, 143-50.

Abdullahi, O., A. Karani, C.C. Tigoi, D. Mugo, S. Kungu, E. Wanjiru, J. Jomo, R. Musyimi, M. Lipsitch and J.A. Scott. (2012). Rates of acquisition and clearance of pneumococcal serotypes in the nasopharynges of children in Kilifi District, Kenya. *J Infect Dis* **206**, 1020-9.

Almeida, S.T., S. Nunes, A.C. Santos Paulo, I. Valadares, S. Martins, F. Breia, A. Brito-Avô, A. Morais, H. de Lencastre and R. Sá-Leão. (2014). Low prevalence of pneumococcal carriage and high serotype and genotype diversity among adults over 60 years of age living in Portugal. *PLoS One* **9**, e90974.

Aniansson, G., B. Alm, B. Andersson, P. Larsson, O. Nylen, H. Peterson, P. Rigner, M. Svanborg and C. Svanborg. (1992). Nasopharyngeal colonization during the first year of life. *J Infect Dis* **165 Suppl 1**, S38-42.

Arbique, J.C., C. Poyart, P. Trieu-Cuot, G. Quesne, G. Carvalho Mda, A.G. Steigerwalt, R.E. Morey, D. Jackson, R.J. Davidson and R.R. Facklam. (2004). Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. *J Clin Microbiol* **42**, 4686-96.

Austrian, R. (1960). The Gram stain and the etiology of lobar pneumonia, an historical note. *Bacteriol Rev* **24**, 261-5.

Austrian, R. (1999). The pneumococcus at the millennium: not down, not out. *J Infect Dis* **179 Suppl 2**, S338-41.

Austrian, R., H.P. Bernheimer, E.E. Smith and G.T. Mills. (1959). Simultaneous production of two capsular polysaccharides by pneumococcus. II. The genetic and biochemical bases of binary capsulation. *J Exp Med* **110**, 585-602.

Austrian, R. and J. Gold. (1964). Pneumococcal Bacteremia with Especial Reference to Bacteremic Pneumococcal Pneumonia. *Ann Intern Med* **60**, 759-76.

Avery, O.T., C.M. Macleod and M. McCarty. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-58.

Beighton, D., A.D. Carr and B.A. Oppenheim. (1994). Identification of viridans streptococci associated with bacteraemia in neutropenic cancer patients. *J Med Microbiol* **40**, 202-4.

Bentley, S.D., D.M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M.A. Quail, G. Samuel, I.C. Skovsted, M.S. Kaltoft, B. Barrell, P.R. Reeves, J. Parkhill and B.G. Spratt. (2006). Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2**, e31.

Bishop, C.J., D.M. Aanensen, G.E. Jordan, M. Kilian, W.P. Hanage and B.G. Spratt. (2009). Assigning strains to bacterial species via the internet. *BMC Biol* **7**, 3.

Bogaert, D., R. De Groot and P.W. Hermans. (2004). *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *Lancet Infect Dis* **4**, 144-54.

Bogaert, D., B. Keijser, S. Huse, J. Rossen, R. Veenhoven, E. van Gils, J. Bruin, R. Montijn, M. Bonten and E. Sanders. (2011). Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PLoS One* **6**, e17035.

Borek, A.P., D.C. Dressel, J. Hussong and L.R. Peterson. (1997). Evolving clinical problems with *Streptococcus pneumoniae*: increasing resistance to antimicrobial agents, and failure of traditional optochin identification in Chicago, Illinois, between 1993 and 1996. *Diagn Microbiol Infect Dis* **29**, 209-14.

Brown, J.S., S.M. Gilliland and D.W. Holden. (2001). A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol* **40**, 572-85.

Buck, J.M., C. Lexau, M. Shapiro, A. Glennen, D.J. Boxrud, B. Koziol, C.G. Whitney, B. Beall, R. Danila and R. Lynfield. (2006). A community outbreak of conjunctivitis caused by nontypeable *Streptococcus pneumoniae* in Minnesota. *Pediatr Infect Dis J* **25**, 906-11.

Carvalho Mda, G., M.L. Tondella, K. McCaustland, L. Weidlich, L. McGee, L.W. Mayer, A. Steigerwalt, M. Whaley, R.R. Facklam, B. Fields, G. Carlone, E.W. Ades, R. Dagan and J.S. Sampson. (2007). Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *J Clin Microbiol* **45**, 2460-6.

CDC. (2000). Preventing pneumococcal disease among infants and young children. Recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep* **49**, 1-35.

CDC. (2013). Antibiotic Resistance Threats in the United States, 2013. *U.S. Department of Health and Human Services, Centers for Disease Control and Prevention*

Chen, J.H., K.K. She, O.Y. Wong, J.L. Teng, W.C. Yam, S.K. Lau, P.C. Woo, V.C. Cheng and K.Y. Yuen. (2015). Use of MALDI Biotyper plus ClinProTools mass spectra analysis for correct identification of *Streptococcus pneumoniae* and *Streptococcus mitis*/*oralis*. *J Clin Pathol* **68**, 652-6.

Chewapreecha, C., S.R. Harris, N.J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D.M. Aanensen, A.E. Mather, A.J. Page, S.J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner and S.D. Bentley. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46**, 305-9.

Dagan, R. and K.P. Klugman. (2008). Impact of conjugate pneumococcal vaccines on antibiotic resistance. *Lancet Infect Dis* **8**, 785-95.

Donati, C., N.L. Hiller, H. Tettelin, A. Muzzi, N.J. Croucher, S.V. Angiuoli, M. Oggioni, J.C. Dunning Hotopp, F.Z. Hu, D.R. Riley, A. Covacci, T.J. Mitchell, S.D. Bentley, M. Kilian, G.D. Ehrlich, R. Rappuoli, E.R. Moxon and V. Masignani. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**, R107.

Donkor, E.S. (2013). Understanding the pneumococcus: transmission and evolution. *Front Cell Infect Microbiol* **3**, 7.

Douglas, C.W., J. Heath, K.K. Hampton and F.E. Preston. (1993). Identity of viridans streptococci isolated from cases of infective endocarditis. *J Med Microbiol* **39**, 179-82.

Drancourt, M. (2010). Detection of microorganisms in blood specimens using matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a review. *Clin Microbiol Infect* **16**, 1620-5.

El Aila, N.A., S. Emler, T. Kaijalainen, T. De Baere, B. Saerens, E. Alkan, P. Deschaght, R. Verhelst and M. Vaneechoutte. (2010). The development of a 16S rRNA gene based PCR for the identification of *Streptococcus pneumoniae* and comparison with four other species specific PCR assays. *BMC Infect Dis* **10**, 104.

Enright, M.C. and B.G. Spratt. (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* **144 (Pt 11)**, 3049-60.

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin Microbiol Rev* **15**, 613-30.

Feldman, C. and R. Anderson. (2014). Review: current and new generation pneumococcal vaccines. *J Infect* **69**, 309-25.

Fenoll, A., R. Munoz, E. Garcia and A.G. de la Campa. (1994). Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the F0 complex of the *Streptococcus pneumoniae* and *Streptococcus oralis* H(+)-ATPases. *Mol Microbiol* **12**, 587-98.

Fenselau, C. and P.A. Demirev. (2001). Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom Rev* **20**, 157-71.

Garcia, E., D. Llull, R. Munoz, M. Mollerach and R. Lopez. (2000). Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. *Res Microbiol* **151**, 429-35.

Geno, K.A., G.L. Gilbert, J.Y. Song, I.C. Skovsted, K.P. Klugman, C. Jones, H.B. Konradsen and M.H. Nahm. (2015). Pneumococcal Capsules and Their Types: Past, Present, and Future. *Clin Microbiol Rev* **28**, 871-99.

Greisen, K., M. Loeffelholz, A. Purohit and D. Leong. (1994). PCR primers and probes for the 16S rRNA gene of most species of pathogenic bacteria, including bacteria found in cerebrospinal fluid. *J Clin Microbiol* **32**, 335-51.

Guidolin, A., J.K. Morona, R. Morona, D. Hansman and J.C. Paton. (1994). Nucleotide sequence analysis of genes essential for capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 19F. *Infect Immun* **62**, 5384-96.

Hakenbeck, R., K. Kaminski, A. Konig, M. van der Linden, J. Paik, P. Reichmann and D. Zahner. (1999). Penicillin-binding proteins in beta-lactam-resistant *Streptococcus pneumoniae*. *Microb Drug Resist* **5**, 91-9.

Hanage, W.P., T. Kaijalainen, E. Herva, A. Saukkoriipi, R. Syrjanen and B.G. Spratt. (2005). Using multilocus sequence data to define the pneumococcus. *J Bacteriol* **187**, 6223-30.

Hanage, W.P., T. Kaijalainen, A. Saukkoriipi, J.L. Rickcord and B.G. Spratt. (2006). A successful, diverse disease-associated lineage of nontypeable pneumococci that has lost the capsular biosynthesis locus. *J Clin Microbiol* **44**, 743-9.

Hardie, J.M. and R.A. Whiley. (1997). Classification and overview of the genera *Streptococcus* and *Enterococcus*. *Soc Appl Bacteriol Symp Ser* **26**, 1S-11S.

Harf-Monteil, C., C. Granello, C. Le Brun, H. Monteil and P. Riegel. (2006). Incidence and pathogenic effect of *Streptococcus pseudopneumoniae*. *J Clin Microbiol* **44**, 2240-1.

Hathaway, L.J., S.D. Brugger, B. Morand, M. Bangert, J.U. Rotzetter, C. Hauser, W.A. Graber, S. Gore, A. Kadioglu and K. Muhlemann. (2012). Capsule type of *Streptococcus pneumoniae* determines growth phenotype. *PLoS Pathog* **8**, e1002574.

Hathaway, L.J., P. Stutzmann Meier, P. Battig, S. Aebi and K. Muhlemann. (2004). A homologue of *aliB* is found in the capsule region of nonencapsulated *Streptococcus pneumoniae*. *J Bacteriol* **186**, 3721-9.

Hausdorff, W.P., J. Bryant, C. Kloek, P.R. Paradiso and G.R. Siber. (2000a). The contribution of specific pneumococcal serogroups to different disease manifestations: implications for conjugate vaccine formulation and use, part II. *Clin Infect Dis* **30**, 122-40.

Hausdorff, W.P., J. Bryant, P.R. Paradiso and G.R. Siber. (2000b). Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I. *Clin Infect Dis* **30**, 100-21.

Heidelberger, M. and O.T. Avery. (1923). The Soluble Specific Substance of Pneumococcus. *J Exp Med* **38**, 73-9.

Henrichsen, J. (1995). Six newly recognized types of *Streptococcus pneumoniae*. *J Clin Microbiol* **33**, 2759-62.

Henriques-Normark, B. and E.I. Tuomanen. (2013). The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harb Perspect Med* **3**,

Hicks, L.A., L.H. Harrison, B. Flannery, J.L. Hadler, W. Schaffner, A.S. Craig, D. Jackson, A. Thomas, B. Beall, R. Lynfield, A. Reingold, M.M. Farley and C.G. Whitney. (2007). Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998-2004. *J Infect Dis* **196**, 1346-54.

Hilty, M., D. Wuthrich, S.J. Salter, H. Engel, S. Campbell, R. Sá-Leão, H. de Lencastre, P. Hermans, E. Sadowy, P. Turner, C. Chewapreecha, M. Diggle, G. Pluschke, L. McGee, O. Koseoglu Eser, D.E. Low, H. Smith-Vaughan, A. Endimiani, M. Kuffer, M. Dupasquier, E. Beaudoing, J. Weber, R. Bruggmann, W.P. Hanage, J. Parkhill, L.J. Hathaway, K. Muhlemann and S.D. Bentley. (2014). Global phylogenomic analysis of nonencapsulated *Streptococcus pneumoniae* reveals a deep-branching classic lineage that is distinct from multiple sporadic lineages. *Genome Biol Evol* **6**, 3281-94.

Hussain, M., A. Melegaro, R.G. Pebody, R. George, W.J. Edmunds, R. Talukdar, S.A. Martin, A. Efstratiou and E. Miller. (2005). A longitudinal household study of *Streptococcus pneumoniae* nasopharyngeal carriage in a UK setting. *Epidemiol Infect* **133**, 891-8.

Hyams, C., E. Camberlein, J.M. Cohen, K. Bax and J.S. Brown. (2010). The *Streptococcus pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. *Infect Immun* **78**, 704-15.

Ikryannikova, L.N., A.V. Filimonova, M.V. Malakhova, T. Savinova, O. Filimonova, E.N. Ilina, V.A. Dubovickaya, S.V. Sidorenko and V.M. Govorun. (2013). Discrimination between *Streptococcus pneumoniae* and *Streptococcus mitis* based on sorting of their MALDI mass spectra. *Clin Microbiol Infect* **19**, 1066-71.

Ikryannikova, L.N., K.N. Lapin, M.V. Malakhova, A.V. Filimonova, E.N. Ilina, V.A. Dubovickaya, S.V. Sidorenko and V.M. Govorun. (2011). Misidentification of alpha-hemolytic streptococci by routine tests in clinical practice. *Infect Genet Evol* **11**, 1709-15.

Jado, I., A. Fenoll, J. Casal and A. Perez. (2001). Identification of the *psaA* gene, coding for pneumococcal surface adhesin A, in viridans group streptococci other than *Streptococcus pneumoniae*. *Clin Diagn Lab Immunol* **8**, 895-8.

Johnsborg, O., V. Eldholm and L.S. Havarstein. (2007). Natural genetic transformation: prevalence, mechanisms and function. *Res Microbiol* **158**, 767-78.

Johnston, C., J. Hinds, A. Smith, M. van der Linden, J. Van Eldere and T.J. Mitchell. (2010). Detection of large numbers of pneumococcal virulence genes in streptococci of the mitis group. *J Clin Microbiol* **48**, 2762-9.

Kaijalainen, T., A. Saukkoriipi, A. Bloigu, E. Herva and M. Leinonen. (2005). Real-time pneumolysin polymerase chain reaction with melting curve analysis differentiates pneumococcus from other alpha-hemolytic streptococci. *Diagn Microbiol Infect Dis* **53**, 293-9.

Kawamura, Y., X.G. Hou, F. Sultana, H. Miura and T. Ezaki. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *Int J Syst Bacteriol* **45**, 406-8.

Keith, E.R., R.G. Podmore, T.P. Anderson and D.R. Murdoch. (2006). Characteristics of *Streptococcus pseudopneumoniae* isolated from purulent sputum samples. *J Clin Microbiol* **44**, 923-7.

Keller, L.E., C.V. Jones, J.A. Thornton, M.E. Sanders, E. Swiatlo, M.H. Nahm, I.H. Park and L.S. McDaniel. (2013). PspK of *Streptococcus pneumoniae* increases adherence to epithelial cells and enhances nasopharyngeal colonization. *Infect Immun* **81**, 173-81.

Kellner, J. (2011). Update on the success of the pneumococcal conjugate vaccine. *Paediatr Child Health* **16**, 233-40.

Kilian, M., K. Poulsen, T. Blomqvist, L.S. Havarstein, M. Bek-Thomsen, H. Tettelin and U.B. Sorensen. (2008). Evolution of *Streptococcus pneumoniae* and its close commensal relatives. *PLoS One* **3**, e2683.

Kim, J.O., S. Romero-Steiner, U.B. Sorensen, J. Blom, M. Carvalho, S. Barnard, G. Carlone and J.N. Weiser. (1999). Relationship between cell surface carbohydrates and intrastrain variation on opsonophagocytosis of *Streptococcus pneumoniae*. *Infect Immun* **67**, 2327-33.

Klugman, K.P. (1990). Pneumococcal resistance to antibiotics. *Clin Microbiol Rev* **3**, 171-96.

Klugman, K.P., S.A. Madhi and W.C. Albrich. (2008). Novel approaches to the identification of *Streptococcus pneumoniae* as the cause of community-acquired pneumonia. *Clin Infect Dis* **47 Suppl 3**, S202-6.

Ko, K.S., J.Y. Baek and J.H. Song. (2013). Capsular gene sequences and genotypes of "serotype 6E" *Streptococcus pneumoniae* isolates. *J Clin Microbiol* **51**, 3395-9.

Leinonen, M., A. Sakkinen, R. Kalliokoski, J. Luotonen, M. Timonen and P.H. Makela. (1986). Antibody response to 14-valent pneumococcal capsular polysaccharide vaccine in pre-school age children. *Pediatr Infect Dis* **5**, 39-44.

Levine, H., S. Zarka, R. Dagan, T. Sela, V. Rozhavski, D.I. Cohen and R.D. Balicer. (2012). Transmission of *Streptococcus pneumoniae* in adults may occur through saliva. *Epidemiol Infect* **140**, 561-5.

Llull, D., R. Lopez and E. Garcia. (2006). Characteristic signatures of the *lytA* gene provide a basis for rapid and reliable diagnosis of *Streptococcus pneumoniae* infections. *J Clin Microbiol* **44**, 1250-6.

Llull, D., R. Munoz, R. Lopez and E. Garcia. (1999). A single gene (*tts*) located outside the *cap* locus directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide. Type 37 pneumococci are natural, genetically binary strains. *J Exp Med* **190**, 241-51.

Lopez, R. (2006). Pneumococcus: the sugar-coated bacteria. *Int Microbiol* **9**, 179-90.

Lucas, V.S., D. Beighton, G.J. Roberts and S.J. Challacombe. (1997). Changes in the oral streptococcal flora of children undergoing allogeneic bone marrow transplantation. *J Infect* **35**, 135-41.

Maiden, M.C., J.A. Bygraves, E. Feil, G. Morelli, J.E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D.A. Caugant, I.M. Feavers, M. Achtman and B.G. Spratt. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* **95**, 3140-5.

Marsh, R., H. Smith-Vaughan, K.M. Hare, M. Binks, F. Kong, J. Warning, G.L. Gilbert, P. Morris and A.J. Leach. (2010). The nonserotypeable pneumococcus: phenotypic dynamics in the era of anticapsular vaccines. *J Clin Microbiol* **48**, 831-5.

Martin, M., J.H. Turco, M.E. Zegans, R.R. Facklam, S. Sodha, J.A. Elliott, J.H. Pryor, B. Beall, D.D. Erdman, Y.Y. Baumgartner, P.A. Sanchez, J.D. Schwartzman, J. Montero, A. Schuchat and C.G.

Whitney. (2003). An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. *N Engl J Med* **348**, 1112-21.

Messmer, T.O., J.S. Sampson, A. Stinson, B. Wong, G.M. Carlone and R.R. Facklam. (2004). Comparison of four polymerase chain reaction assays for specificity in the identification of *Streptococcus pneumoniae*. *Diagn Microbiol Infect Dis* **49**, 249-54.

Morales, M., A.J. Martin-Galiano, M. Domenech and E. Garcia. (2015). Insights into the Evolutionary Relationships of LytA Autolysin and Ply Pneumolysin-Like Genes in *Streptococcus pneumoniae* and Related Streptococci. *Genome Biol Evol* **7**, 2747-61.

Mundy, L.S., E.N. Janoff, K.E. Schwebke, C.J. Shanholtzer and K.E. Willard. (1998). Ambiguity in the identification of *Streptococcus pneumoniae*. Optochin, bile solubility, quellung, and the AccuProbe DNA probe tests. *Am J Clin Pathol* **109**, 55-61.

Munoz, R., A. Fenoll, D. Vicioso and J. Casal. (1990). Optochin-resistant variants of *Streptococcus pneumoniae*. *Diagn Microbiol Infect Dis* **13**, 63-6.

Murray, P.R., K.S. Rosenthal and M.A. Pfaller, 2005. *Streptococcus*. In: Medical Microbiology, chapter 23, pages 237-258.

Nakajima, T., S. Nakanishi, C. Mason, J. Montgomery, P. Leggett, M. Matsuda, W.A. Coulter, B.C. Millar, C.E. Goldsmith and J.E. Moore. (2013). Population structure and characterization of viridans group streptococci (VGS) isolated from the upper respiratory tract of patients in the community. *Ulster Med J* **82**, 164-8.

Neeleman, C., C.H. Klaassen, D.M. Klomberg, H.A. de Valk and J.W. Mouton. (2004). Pneumolysin is a key factor in misidentification of macrolide-resistant *Streptococcus pneumoniae* and is a putative virulence factor of *S. mitis* and other streptococci. *J Clin Microbiol* **42**, 4355-7.

Nelson, A.L., A.M. Roche, J.M. Gould, K. Chim, A.J. Ratner and J.N. Weiser. (2007). Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. *Infect Immun* **75**, 83-90.

Nunes, S., R. Sá-Leão, J. Carriço, C.R. Alves, R. Mato, A.B. Avo, J. Saldanha, J.S. Almeida, I.S. Sanches and H. de Lencastre. (2005). Trends in drug resistance, serotypes, and molecular types of *Streptococcus pneumoniae* colonizing preschool-age children attending day care centers in Lisbon, Portugal: a summary of 4 years of annual surveillance. *J Clin Microbiol* **43**, 1285-93.

Nunes, S., R. Sá-Leão and H. de Lencastre. (2008). Optochin resistance among *Streptococcus pneumoniae* strains colonizing healthy children in Portugal. *J Clin Microbiol* **46**, 321-4.

O'Brien, K.L., M.C. Steinhoff, K. Edwards, H. Keyserling, M.L. Thoms and D. Madore. (1996). Immunologic priming of young children by pneumococcal glycoprotein conjugate, but not polysaccharide, vaccines. *Pediatr Infect Dis J* **15**, 425-30.

O'Brien, K.L., L.J. Wolfson, J.P. Watt, E. Henkle, M. Deloria-Knoll, N. McCall, E. Lee, K. Mulholland, O.S. Levine and T. Cherian. (2009). Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet* **374**, 893-902.

Obregon, V., P. Garcia, E. Garcia, A. Fenoll, R. Lopez and J.L. Garcia. (2002). Molecular peculiarities of the *lytA* gene isolated from clinical pneumococcal strains that are bile insoluble. *J Clin Microbiol* **40**, 2545-54.

Park, I.H., K.A. Geno, L.K. Sherwood, M.H. Nahm and B. Beall. (2014). Population-based analysis of invasive nontypeable pneumococci reveals that most have defective capsule synthesis genes. *PLoS One* **9**, e97825.

Park, I.H., K.H. Kim, A.L. Andrade, D.E. Briles, L.S. McDaniel and M.H. Nahm. (2012). Nontypeable pneumococci can be divided into multiple *cps* types, including one type expressing the novel gene *pspK*. *MBio* **3**,

Phillips, G., R. Barker and O. Brogan. (1988). Optochin-resistant *Streptococcus pneumoniae*. *Lancet* **2**, 281.

Pikis, A., J.M. Campos, W.J. Rodriguez and J.M. Keith. (2001). Optochin resistance in *Streptococcus pneumoniae*: mechanism, significance, and clinical implications. *J Infect Dis* **184**, 582-90.

Pozzi, G., M.R. Oggioni and A. Tomasz. (1989). DNA probe for identification of *Streptococcus pneumoniae*. *J Clin Microbiol* **27**, 370-2.

Regev-Yochay, G., M. Raz, R. Dagan, N. Porat, B. Shainberg, E. Pinco, N. Keller and E. Rubinstein. (2004). Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. *Clin Infect Dis* **38**, 632-9.

Richter, S.S., K.P. Heilmann, C.L. Dohrn, F. Riahi, S.E. Beekmann and G.V. Doern. (2008). Accuracy of phenotypic methods for identification of *Streptococcus pneumoniae* isolates included in surveillance programs. *J Clin Microbiol* **46**, 2184-8.

Richter, S.S., K.P. Heilmann, C.L. Dohrn, F. Riahi, S.E. Beekmann and G.V. Doern. (2009). Changing epidemiology of antimicrobial-resistant *Streptococcus pneumoniae* in the United States, 2004-2005. *Clin Infect Dis* **48**, e23-33.

Robbins, J.B., R. Austrian, C.J. Lee, S.C. Rastogi, G. Schiffman, J. Henrichsen, P.H. Makela, C.V. Broome, R.R. Facklam, R.H. Tiesjema and et al. (1983). Considerations for formulating the second-generation pneumococcal capsular polysaccharide vaccine with emphasis on the cross-reactive types within groups. *J Infect Dis* **148**, 1136-59.

Rolo, D., S.S. A, A. Domenech, A. Fenoll, J. Linares, H. de Lencastre, C. Ardanuy and R. Sá-Leão. (2013). Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. *PLoS One* **8**, e57047.

Romero, P., R. Lopez and E. Garcia. (2004). Characterization of LytA-like N-acetylmuramoyl-L-alanine amidases from two new *Streptococcus mitis* bacteriophages provides insights into the properties of the major pneumococcal autolysin. *J Bacteriol* **186**, 8229-39.

Russell, N. (1988). Oswald Avery and the origin of molecular biology. *Br J Hist Sci* **21**, 193-400.

Sá-Leão, R., S. Nunes, A. Brito-Avô, C.R. Alves, J.A. Carriço, J. Saldanha, J.S. Almeida, I. Santos-Sanches and H. de Lencastre. (2008). High rates of transmission of and colonization by *Streptococcus pneumonia*e and *Haemophilus influenzae* within a day care center revealed in a longitudinal study. *J Clin Microbiol* **46**, 225-34.

Sá-Leão, R., S. Nunes, A. Brito-Avô, N. Frazão, A.S. Simões, M.I. Crisóstomo, A.C. Paulo, J. Saldanha, I. Santos-Sanches and H. de Lencastre. (2009). Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. *Clin Microbiol Infect* **15**, 1002-7.

Sá-Leão, R., A.S. Simões, S. Nunes, N.G. Sousa, N. Frazão and H. de Lencastre. (2006). Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. *Microbiology* **152**, 367-76.

Sá-Leão, R., A. Tomasz, I.S. Sanches, S. Nunes, C.R. Alves, A.B. Avô, J. Saldanha, K.G. Kristinsson and H. de Lencastre. (2000). Genetic diversity and clonal patterns among antibiotic-susceptible and -resistant *Streptococcus pneumoniae* colonizing children: day care centers as autonomous epidemiological units. *J Clin Microbiol* **38**, 4137-44.

Sabat, A.J., A. Budimir, D. Nashev, R. Sá-Leão, J. van Dijl, F. Laurent, H. Grundmann, A.W. Friedrich and E.S.G.o.E. Markers. (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill* **18**, 20380.

Sakamoto, M., Y. Huang, M. Ohnishi, M. Umeda, I. Ishikawa and Y. Benno. (2004). Changes in oral microbial profiles after periodontal treatment as determined by molecular analysis of 16S rRNA genes. *J Med Microbiol* **53**, 563-71.

Salter, S.J., J. Hinds, K.A. Gould, L. Lambertsen, W.P. Hanage, M. Antonio, P. Turner, P.W. Hermans, H.J. Bootsma, K.L. O'Brien and S.D. Bentley. (2012). Variation at the capsule locus, *cps*, of mistyped and non-typable *Streptococcus pneumoniae* isolates. *Microbiology* **158**, 1560-9.

Satzke, C., P. Turner, A. Virolainen-Julkunen, P.V. Adrian, M. Antonio, K.M. Hare, A.M. Henao-Restrepo, A.J. Leach, K.P. Klugman, B.D. Porter, R. Sa-Leao, J.A. Scott, H. Nohynek, K.L. O'Brien and W.H.O.P.C.W. Group. (2013). Standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*: updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. *Vaccine* **32**, 165-79.

Scholz, C.F., K. Poulsen and M. Kilian. (2012). Novel molecular method for identification of *Streptococcus pneumoniae* applicable to clinical microbiology and 16S rRNA sequence-based microbiome studies. *J Clin Microbiol* **50**, 1968-73.

Scott, J.R., J. Hinds, K.A. Gould, E.V. Millar, R. Reid, M. Santosham, K.L. O'Brien and W.P. Hanage. (2012). Nontypeable pneumococcal isolates among navajo and white mountain apache communities: are these really a cause of invasive disease? *J Infect Dis* **206**, 73-80.

Shelburne, S.A., P. Sahasrabhojane, M. Saldana, H. Yao, X. Su, N. Horstmann, E. Thompson and A.R. Flores. (2014). *Streptococcus mitis* strains causing severe clinical disease in cancer patients. *Emerg Infect Dis* **20**, 762-71.

Simell, B., K. Auranen, H. Kayhty, D. Goldblatt, R. Dagan, K.L. O'Brien and G. Pneumococcal Carriage. (2012). The fundamental link between pneumococcal carriage and disease. *Expert Rev Vaccines* **11**, 841-55.

Simões, A.S., L. Pereira, S. Nunes, A. Brito-Avo, H. de Lencastre and R. Sá-Leão. (2011a). Clonal evolution leading to maintenance of antibiotic resistance rates among colonizing Pneumococci in the PCV7 era in Portugal. *J Clin Microbiol* **49**, 2810-7.

Simões, A.S., R. Sá-Leão, M.J. Eleveld, D.A. Tavares, J.A. Carriço, H.J. Bootsma and P.W. Hermans. (2010). Highly penicillin-resistant multidrug-resistant pneumococcus-like strains colonizing children in Oeiras, Portugal: genomic characteristics and implications for surveillance. *J Clin Microbiol* **48**, 238-46.

Simões, A.S., C. Valente, H. de Lencastre and R. Sá-Leão. (2011b). Rapid identification of noncapsulated *Streptococcus pneumoniae* in nasopharyngeal samples allowing detection of co-colonization and reevaluation of prevalence. *Diagn Microbiol Infect Dis* **71**, 208-16.

Song, J.H., R. Dagan, K.P. Klugman and B. Fritzell. (2012). The relationship between pneumococcal serotypes and antibiotic resistance. *Vaccine* **30**, 2728-37.

Sorensen, U.B., J. Henrichsen, H.C. Chen and S.C. Szu. (1990). Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods. *Microb Pathog* **8**, 325-34.

Suau, A., R. Bonnet, M. Sutren, J.J. Godon, G.R. Gibson, M.D. Collins and J. Dore. (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65**, 4799-807.

Tomasz, A. (1965). Control of the competent state in Pneumococcus by a hormone-like cell product: an example for a new type of regulatory mechanism in bacteria. *Nature* **208**, 155-9.

Trzcinski, K., D. Bogaert, A. Wyllie, M.L. Chu, A. van der Ende, J.P. Bruin, G. van den Dobbelsteen, R.H. Veenhoven and E.A. Sanders. (2013). Superiority of trans-oral over trans-nasal sampling in detecting *Streptococcus pneumoniae* colonization in adults. *PLoS One* **8**, e60520.

Tunkel, A.R. and K.A. Sepkowitz. (2002). Infections caused by viridans streptococci in patients with neutropenia. *Clin Infect Dis* **34**, 1524-9.

Valente, C., J. Hinds, F. Pinto, S.D. Brugger, K. Gould, K. Muhlemann, H. de Lencastre and R. Sá-Leão. (2012). Decrease in pneumococcal co-colonization following vaccination with the seven-valent pneumococcal conjugate vaccine. *PLoS One* **7**, e30235.

van Belkum, A., M. Welker, M. Erhard and S. Chatellier. (2012). Biomedical mass spectrometry in today's and tomorrow's clinical microbiology laboratories. *J Clin Microbiol* **50**, 1513-7.

van Kuppeveld, F.J., J.T. van der Logt, A.F. Angulo, M.J. van Zoest, W.G. Quint, H.G. Niesters, J.M. Galama and W.J. Melchers. (1992). Genus- and species-specific identification of mycoplasmas by 16S rRNA amplification. *Appl Environ Microbiol* **58**, 2606-15.

Walsh, R.L. and A. Camilli. (2011). *Streptococcus pneumoniae* is desiccation tolerant and infectious upon rehydration. *MBio* **2**, e00092-11.

Watson, D.A., D.M. Musher, J.W. Jacobson and J. Verhoef. (1993). A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery. *Clin Infect Dis* **17**, 913-24.

Werno, A.M., M. Christner, T.P. Anderson and D.R. Murdoch. (2012). Differentiation of *Streptococcus pneumoniae* from nonpneumococcal streptococci of the *Streptococcus mitis* group by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J Clin Microbiol* **50**, 2863-7.

Werno, A.M. and D.R. Murdoch. (2008). Medical microbiology: laboratory diagnosis of invasive pneumococcal disease. *Clin Infect Dis* **46**, 926-32.

Wester, C.W., D. Ariga, C. Nathan, T.W. Rice, J. Pulvirenti, R. Patel, F. Kocka, J. Ortiz and R.A. Weinstein. (2002). Possible overestimation of penicillin resistant *Streptococcus pneumoniae* colonization rates due to misidentification of oropharyngeal streptococci. *Diagn Microbiol Infect Dis* **42**, 263-8.

Whalan, R.H., S.G. Funnell, L.D. Bowler, M.J. Hudson, A. Robinson and C.G. Dowson. (2006). Distribution and genetic diversity of the ABC transporter lipoproteins PiuA and PiaA within *Streptococcus pneumoniae* and related streptococci. *J Bacteriol* **188**, 1031-8.

Whatmore, A.M., A. Efstratiou, A.P. Pickerill, K. Broughton, G. Woodard, D. Sturgeon, R. George and C.G. Dowson. (2000). Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "Atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. *Infect Immun* **68**, 1374-82.

Whitney, C.G., M.M. Farley, J. Hadler, L.H. Harrison, N.M. Bennett, R. Lynfield, A. Reingold, P.R. Cieslak, T. Pilishvili, D. Jackson, R.R. Facklam, J.H. Jorgensen, A. Schuchat and N. Active Bacterial Core Surveillance of the Emerging Infections Program. (2003). Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine. *N Engl J Med* **348**, 1737-46.

Wyllie, A.L., M.L. Chu, M.H. Schellens, J. van Engelsdorp Gastelaars, M.D. Jansen, A. van der Ende, D. Bogaert, E.A. Sanders and K. Trzcinski. (2014). *Streptococcus pneumoniae* in saliva of Dutch primary school children. *PLoS One* **9**, e102045.

Yahiaoui, R.Y., C. den Heijer, P. Wolfs, C.A. Bruggeman and E.E. Stobberingh. (2016). Evaluation of phenotypic and molecular methods for identification of *Streptococcus pneumoniae*. *Future Microbiol* **11**, 43-50.

Chapter 2

Non-typeable pneumococci circulating in Portugal are of *cps* type NCC2 and have genomic features typical of encapsulated isolates

Published in: D. A. Tavares[*], A. S. Simões[*], H. J. Bootsma, P. W. M. Hermans, H. de Lencastre, and R. Sá-Leão (2014) *BMC Genomics* **15:** 863-77. *Equal contribution.

Contributions:

D. A Tavares was responsible for all experimental work with the exception of the initial selection of the strains, pulsed-field gel electrophoresis (PFGE), and prophage detection by southern hybridisation of PFGE restriction profiles with a *lytA* probe, which was performed by A. S. Simões.

Summary

Pneumococcus is a major human pathogen and the polysaccharide capsule is considered its main virulence factor. Nevertheless, strains lacking a capsule, named non-typeable pneumococcus (NT), are maintained in nature and frequently colonise the human nasopharynx. Interest in these strains, not targeted by any of the currently available pneumococcal vaccines, has been rising as they seem to play an important role in the evolution of the species. Currently, there is a paucity of data regarding this group of pneumococci. Also, questions have been raised on whether they are true pneumococci. We aimed to obtain insights in the genetic content of NT and the mechanisms leading to non-typeability and to genetic diversity.

A collection of 52 NT isolates representative of the lineages circulating in Portugal between 1997 and 2007, as determined by pulsed-field gel electrophoresis and multilocus sequence typing, was analysed. The capsular region was sequenced and comparative genomic hybridisation (CGH) using a microarray covering the genome of 10 pneumococcal strains was carried out. The presence of mobile elements was investigated as source of intraclonal variation. NT circulating in Portugal were found to have similar capsular regions, of *cps* type NCC2, i.e., having *aliB*-like ORF1 and *aliB*-like ORF2 genes. The core genome of NT was essentially similar to that of encapsulated strains. Also, competence genes and most virulence genes were present. The few virulence genes absent in all NT were the capsular genes, type-I and type-II pili, choline-binding protein A (*cbpA*/*pspC*), and pneumococcal surface protein A (*pspA*). Intraclonal variation could not be entirely explained by the presence of prophages and other mobile elements.

NT circulating in Portugal are a homogeneous group belonging to *cps* type NCC2. Our observations support the theory that they are *bona-fide* pneumococcal isolates that do not express the capsule but are otherwise essentially similar to encapsulated pneumococci. Thus we propose that NT should be routinely identified and reported in surveillance studies.

Introduction

Pneumococcus is a major human pathogen, causing a wide range of infections from otitis media to bacteraemia and meningitis. Its main virulence determinant is a polysaccharide capsule that surrounds pneumococcal cells, providing protection against phagocytosis (Bentley *et al.*, 2006). Together with colony morphology, susceptibility to optochin, and bile solubility, assignment of a serotype (based on the capsular type) has been traditionally the ultimate assay to identify pneumococcus (Kellogg *et al.*, 2001). To date, more than 95 serotypes have been described and, with the exception of type 37, the genes responsible for the expression of the capsule are located in the chromosome between the *dexB* and *aliA* genes (capsular region) (Bentley *et al.*, 2006; Oliver *et al.*, 2013). The pneumococcal capsule is also the target of all currently available pneumococcal vaccines (Nuorti and Whitney, 2010).

Pneumococci lacking a polysaccharide capsule are known to exist in nature and are frequent inhabitants of the upper respiratory tract of humans (Sá-Leão *et al.*, 2006). Although these isolates, often named non-typeable pneumococcus (NT), are mostly

asymptomatically carried in the nasopharynx, they have also been associated with conjunctivitis outbreaks and sporadically associated with other disease manifestations including invasive disease (Martin *et al.*, 2003; Xu *et al.*, 2011; Scott *et al.*, 2012; Rolo *et al.*, 2013). Studies have suggested, using a combination of phenotypic and genotypic methods, that some of these isolates are *bona-fide* pneumococci and share common properties with encapsulated pneumococci (Hanage *et al.*, 2005; Sá-Leão *et al.*, 2006). Also, in vitro studies with non-encapsulated pneumococci have shown that these strains display increased adherence to epithelial tissue, increased capacity for biofilm formation, and are highly transformable (Weiser and Kapoor, 1999; Magee and Yother, 2001; Domenech *et al.*, 2009). Hence, high carriage rates combined with high transformability rates may provide NT with the features needed to play an important role in the evolution of pneumococcus as recently proposed by Chewapreecha *et al.* (Chewapreecha *et al.*, 2014).

In a previous study, we have described the population structure of NT strains in Portugal and identified major lineages associated with them (Sá-Leão *et al.*, 2006). In parallel, others have identified the same lineages in circulation in other geographical settings and the capsular region of NT has been characterised (Hathaway *et al.*, 2004; Hanage *et al.*, 2005; Park *et al.*, 2012; Salter *et al.*, 2012). Based on the capsular region, NT have been proposed to be divided in two groups: Group I includes isolates with a disrupted or non-functional capsular locus and Group II includes isolates with genes not found in conventional capsular types (Hanage *et al.*, 2005). Group II NT have been proposed to be further divided into *cps* types NCC1,

when isolates have the *pspK* gene (pneumococcal surface protein Korea, also referred to as *nspA*, non-typeable pneumococcal surface protein A), encoding for a novel pneumococcal surface protein with several features suggesting a role in cell adhesion and enhanced colonisation, and NCC2, when isolates have both the *aliB*-like ORF1 and *aliB*-like ORF2 genes, predicted to encode for lipoproteins (Hathaway *et al.*, 2004; Park *et al.*, 2012; Salter *et al.*, 2012; Keller *et al.*, 2013). A *cps* type NCC3 has also been described for isolates with *aliB*-like ORF2 but not *aliB*-like ORF1, but these were shown not to be pneumococci (Park *et al.*, 2012).

The observation that several distinct clonal lineages lacking the capsule operon have been in circulation for decades and are not derived from encapsulated strains has raised the question of how different is the genome of these strains compared to encapsulated pneumococci (Martin *et al.*, 2003; Sá-Leão *et al.*, 2006). The aim of this study was to characterise a carriage collection of NT circulating in Portugal in a period of 11 years to obtain insights into the genetic basis of non-typeability and their genomic content and diversity.

Materials and methods

Ethics statement. Approval for the original studies (Sá-Leão *et al.*, 2006; Simões *et al.*, 2011a; Simões *et al.*, 2011b) was obtained from the Ministry of Education. The studies registered and approved at the Health Care Centre of Oeiras that reports to Administração Regional de Saúde (ARS; “Regional Health Administration”) of Lisboa and Vale do Tejo from the Ministry of Health. Signed informed consent was obtained

from parents/guardians of participating children. All samples were coded numerically upon collection and processed anonymously. In the present study, only bacterial isolates were characterised (no human subjects, human material or human data were used). Thus, ethical approval was not required.

Study collection. We selected 52 NT strains for detailed characterisation. This collection was extracted from a total of 422 NT strains isolated between 1997 and 2007 from the nasopharynx of preschool children attending day-care centres in Lisbon, Portugal. The isolates were previously characterised by PFGE, MLST, and antibiotic susceptibility to penicillin, amoxicillin, ceftriaxone, erythromycin, clindamycin, tetracycline, chloramphenicol, and trimethoprim sulfamethoxazole (SXT) (Sá-Leão *et al.*, 2006; Simões *et al.*, 2011a; Simões *et al.*, 2011b). The 52 strains characterised in this study were selected to cover the diversity of profiles observed among the 422 isolates, as determined by PFGE, MLST, and antibiotyping. CCs were defined based on goeBURST classification (goeburst.phyloviz.net).

DNA extraction. Total genomic DNA was isolated using either the DNeasy Blood & Tissue kit (Qiagen, Hilden, Germany), or the High Pure PCR Template Preparation kit (Roche Diagnostics GmbH, Mannheim, Germany), according to the manufacturer's recommendations.

Characterisation of the capsular (*dexB-aliA*) region. The *dexB-aliA* region, corresponding to the capsular region in encapsulated pneumococci, was amplified by PCR using the primers described by Kilian *et al.* using the following conditions: 92°C for 2 min; 30 cycles of 92°C for 10 sec, 58°C for 30 sec, and 68°C for 15 min; and a final extension at 68°C for 7min (Kilian *et al.*, 2008). For a final volume of 50µL, the

PCR mixture contained 20ng of DNA, 1x Expand Long Template buffer 3 with 2.75mM $MgCl_2$ (Roche), 3.2mM (each) deoxynucleoside triphosphates, 0.4mM of each primer, and 3.75U of Expand Long Template enzyme mix (Roche). Amplicons were purified using ExoSAP by incubating 30µL of the PCR product with 6U of Exonuclease I (New England Biolabs, Ipswich, MA, USA) and 6U of Shrimp Alkaline Phosphatase (GE Healthcare, Waukesha, WI, USA) for 30 min at 37°C followed by 15 min at 80°C.

RFLP signatures of the capsular region were determined after digestion of 15µL of purified PCR fragments with HinfI or StyI for 3h at 37°C. For a total volume of 20µL, 5U of enzyme, 1x NEBuffer (New England Biolabs), and 2µg of BSA (for StyI) were added. Results were analysed by gel electrophoresis and Bionumerics software (version 3.0, Applied Maths, Gent, Belgium). Patterns were clustered by UPGMA and a dendrogram was generated from a similarity matrix calculated using the Dice similarity coefficient with an optimisation of 0.5% and a tolerance of 1.0%. RFLP patterns determined by digestion with HinfI were arbitrarily named A to H.

Sequencing of the capsular region of representative RFLP patterns was performed by primer walking. Primers were designed using the nucleotide sequence of strain 110.58 as a template [GenBank:AY653211.1] (Additional file 1) (Hathaway *et al.*, 2004; Kilian *et al.*, 2008). PCR products were obtained, purified, and sent to Macrogen, Inc. (Seoul, South Korea) for sequencing. Additional primers were designed to amplify and sequence the gaps between fragments as needed. Sequences were analysed and aligned using the Lasergene software (DNASTAR Inc., Madison, WI, USA). Nucleotide sequences of the capsular region were further

analysed by performing a nucleotide BLAST search at the National Center for Biotechnology Information Website (blast.ncbi.nlm. nih.gov/Blast.cgi) against the nucleotide database and also against the capsular region sequences previously described for NT strains (Hathaway *et al.*, 2004; Park *et al.*, 2012; Salter *et al.*, 2012).

CGH. Microarrays used in this study were 12x135K NimbleGen arrays (Roche). Labelling, hybridisation, and washing of the samples was done as recommended by the manufacturer using a NimbleGen microarray workflow (Roche): 1µg of DNA from each strain was fluorescently labelled with Cy3 Random Nonamers using the NimbleGen One-Color DNA Labeling kit, samples were hybridised to the microarray slide using the NimbleGen Hybridization System, slides were washed using the NimbleGen Wash Buffer kit, and CGH data was acquired on a NimbleGen MS 200 Scanner. Normalisation and background correction of data was done by quantile RMA analysis using the ArrayStar software (DNASTAR). A cut-off of 512 was reached by drawing a graph of frequencies of signal intensities for all strains. Genes with signal intensities of 512 or above were considered present (assigned 1) and genes with signal intensities bellow that value were considered absent (assigned -1) from a given strain.

Validation of the microarray. The microarray used was designed based on the genome sequence of 10 pneumococcal strains: TIGR4, R6, D39, BHN100, CBR206, LGST215, BHN191, BHN418, Sp14-BS69, and Sp3-BS71 (Baltz *et al.*, 1998; Iannelli *et al.*, 1999; Tettelin *et al.*, 2001; Hiller *et al.*, 2007; Sá-Leão *et al.*, 2008; Rodrigues *et al.*, 2009; Hyams *et al.*, 2011; Browall *et al.*, 2014). Triplicates of probes representing genes present in these strains were added sequentially resulting in 3,052 non-

redundant ORFs. Nine of the 10 strains represented in the array were hybridised with it for validation. Only 16 of 3,052 (0.52%) ORFs present in the microarray gave false negative results (Additional file 2). Most of these genes encoded for hypothetical proteins or mobile elements that might have been lost (during repeated handling). None of the 16 genes were part of the core genome, were related to virulence or located in ARs.

ARs. The presence of ARs (or regions of diversity) previously identified (reviewed in Blomberg *et al.*, 2009) was investigated for NT strains. New ARs were identified as defined by Tettellin and Hollingshead: three or more contiguous genes in the TIGR4 genome that were absent from at least one of the analysed strains (Tettelin and Hollingshead, 2004). Classification of new ARs followed the nomenclature proposed by Blomberg *et al*. and was done sequentially (Blomberg *et al.*, 2009).

Detection and characterisation of genes by PCR. The presence of genes *comC*, *comD*, and *piaA* and the presence of type-I and type-II pili was assessed by PCR and characterised by sequencing when needed. *ComD* was amplified using primers comD_F (ATTAAAGGTGGGGAGATGAGG) and comD_R (CCAGCATAATCATGTCG), designed with TIGR4 [GenBank:NC_003028.3) and R6 [GenBank:NC_003098.1] nucleotide sequences as templates. Amplicons with an expected size of 841bp were amplified using the following conditions: 94°C for 4 min; 30 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 1 min; and a final extension at 72°C for 4 min. For a final volume of 50µL, the PCR mixture contained 1µL DNA, 1x Colorless GoTaq Flexi buffer (Promega, Madison, WI, USA), 2.5mM $MgCl_2$, 80µM (each) deoxynucleoside triphosphates, 0.4mM of each primer, and 2.5U of GoTaq DNA polymerase.

Amplicons were purified using ExoSAP as described above, sent to Macrogen for sequencing, and analysed by using Lasergene software. The presence of *comC* was assessed as described by Whatmore *et al*. or Carrolo *et al.* (Whatmore *et al.*, 1999; Carrolo *et al.*, 2009); the presence of *piaA* was assessed as described by Whalan *et al.* (Whalan *et al.*, 2006), and the presence of type-I and type-II pili as described by Zahner *et al.* (Zahner *et al.*, 2010).

Prophage detection by southern hybridisation of PFGE restriction profiles with a *lytA* probe. Preparation of chromosomal DNA, digestion with SmaI endonuclease, and separation of DNA fragments by PFGE were carried out as previously described (Sá-Leão *et al.*, 2000). Southern blotting of PFGE gels with a probe for the *lytA* gene was performed as previously described (Severina *et al.*, 1999).

Availability of supporting data. Availability of supporting data Microarray data supporting the results of this article have been submitted to NCBI Gene Expression Omnibus (GEO) archive repository (www.ncbi.nlm.nih.gov/geo/). The GEO Series Accession Number is GSE58329.

Results

Capsular region of NT. To obtain insights into the genetic basis of non-typeability, the capsular region was characterised for a set of 42 NT strains representative of the lineages detected in cross-sectional colonisation studies conducted in Portugal among children between 1997 and 2007 (Table 1). Amplification of this region yielded, in all strains, a fragment of 6,000-8,500 bp. To investigate the heterogeneity

Table 1. Study collection and characteristics of the strains.

CC[a]	Strain	Year	PFGE	MLST	Antibiotype (non susceptible to)[b]	CSP/ComD[c]	capsular region RFLP	capsular region sequenced	Analysis by CGH
344	PT944	2001	NT1	344	PG, Ery, Da, Tet, SXT	2/2	A	Yes	Yes
	LGST142	2000	NT1	344	Ery, Da, Tet, SXT	2/2	A	No	No
	PT191	2001	NT1	344	PG, Ery, Da, Tet, SXT	2/2	A	No	No
	PT3412b	2002	NT1	344	Ery, Da, Tet, SXT	2/nd	A	No	No
	PT998	2001	NT1	344	PG, Ery, Tet, SXT	2/nd	A	No	No
	LGST214	2000	ND	344	Ery, Tet, SXT	2/nd	A	No	No
	DCC2367	1999	NT1	344	PG, Tet, SXT	2/2	F	Yes	No
	PT389	2001	NT1	344	PG, Tet, SXT	2/nd	F	No	No
	PT4427a	2002	NT1	344	PG, Ery, Tet, SXT	2/nd	H	Yes	No
	WL212	2001	NT1	1619	PG, Ery, Da, SXT	2/2	A	No	Yes
	PT5899	2007	NT1	5220	PG, Ery, Tet, SXT	2/nd	nd	No	Yes
	DCC635	1997	NT2	344	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	WL992	2002	NT3	344	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	PT2987	2002	NT4	344	PG, Ery, Da, Tet, SXT	2/nd	A	No	No
	PT2293b	2001	NT4	344	PG, Ery, Da, Tet, SXT	2/2	E	Yes	Yes
	PT6317	2007	NT5	344	PG, Ery, Da, Tet, SXT	2/nd	nd	No	Yes
	PT5838b	2007	NT6	344	Ery, Da, Tet, SXT	2/nd	nd	No	Yes
	WL1514	2003	NT7	344	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	PT6318	2007	NT7	4586	PG, Ery, Da, Tet, SXT	2/nd	nd	No	Yes
	PT5269	2006	NT8	344	PG, Ery, Da, Tet, SXT	2/nd	nd	No	Yes
	DCC2879	1999	NT9	897	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	PT1571b	2001	NT10	344	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	PT5727	2006	NT11	344	PG, Ery, Da, Tet, SXT	2/nd	nd	No	Yes
	PT5082a	2003	NT22	344	PG, Ery, Da, Tet, SXT	2/nd	I	No	No
	WL598	2001	NT25	344	PG, Tet, SXT	2/nd	F	No	No
	DCC1795	1998	NT26	1541	PG, Ery, Da, Tet, SXT	2/2	A	No	Yes
	DCC2435p	1999	ND	344	Ery, Da, Tet, SXT	2/nd	A	No	No
1156	PT268	2001	NT21	1156	PG, Ery, Da, Tet, SXT	1/nd	A	No	No
	PT6210	2007	NT21	4583	PG, Ery, Da, Tet, SXT	1/nd	nd	No	Yes
	PT2687b	2001	NT22	1156	PG, Ery, Da, Tet, SXT	1/nd	A	No	No
	PT5561	2006	NT22	1156	PG, Ery, Da, Tet, SXT	1/nd	nd	No	Yes
	PT4014	2002	NT22	1153	PG, Ery, Da, Tet, SXT	1/1	C	No	Yes
	PT4222	2002	NT24	1156	PG, Ery, Da, Tet, SXT	1/1	A	No	No
	PT5002	2003	NT24	1156	Ery, Da, Tet	1/1	A	No	Yes
	PT1493	2001	NT24	1617	PG, SXT	1/1	A	Yes	Yes
	WL352.1	2001	NT24	1703	PG, SXT	1/1	A	No	Yes
	PT3201	2002	NT24	1153	PG, Ery, Da, Tet, SXT	1/1	C	Yes	Yes
	PT6209b	2007	NT24	4583	PG, Ery, Da, Tet	1/nd	nd	No	Yes
	PT2322	2001	ND	1153	PG, Ery, Da, Tet, SXT	1/nd	C	No	No
320	PT1804b	2001	NT19	888	PG, SXT	1/1	A	Yes	Yes
1540*	PT1718	2001	NT12	1540	SXT	1/4	A	Yes	Yes
1278*	PT4812	2003	NT22	1278	PG, SXT	1/1	A	Yes	Yes
941	DCC2787	1999	NT13	941	SXT	2/2	B	Yes	Yes
	WL165b	2001	NT13	1704		2/2	B	No	Yes
	DCC2648	1999	NT14	941	SXT	2/2	B	No	Yes
448	WL850a	2002	NT15	448		2/2	B	Yes	Yes
	WL1084	2002	NT15	448		2/2	B	No	No
	PT2417	2001	NT15	448	PG, SXT	2/nd	B	No	No
	WL108	2001	NT16	448		2/nd	nd	No	Yes
1618	PT673	2001	NT17	1618	PG, Ery	1/1	D	Yes	Yes
	WL402.1b	2001	NT17	1618	PG, Ery, Da, Tet, SXT	1/1	D	No	Yes
1705*	WL977	2002	NT23	1705	PG, SXT	1/1	G	Yes	Yes

a – clonal complex (CC); singleton (*); b – penicillin G (PG), erythromycin (Ery), clindamycin (Da), tetracycline (Tet), and trimethoprim sulfamethoxazole (SXT); c – ComD2 had an E151K substitution and ComD4 had an M77I and an E151K substitutions, both outside the sensor domain of ComD; nd – not determined.

of the capsular region, restriction fragment length polymorphism (RFLP) patterns were determined by digestion with HinfI. Nine different patterns could be distinguished after digestion with HinfI (Figure 1, Table 1). We then selected 13

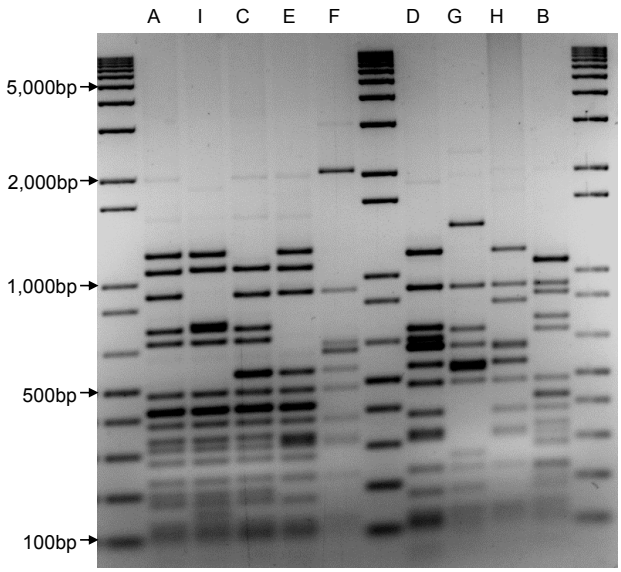


Figure 1. RFLP patterns of the capsular region of NT strains with HinfI. Capital letters in lanes refer to an arbitrary pattern designation.

isolates, representative of the different capsular RFLP patterns found in each CC, for sequencing. The findings are summarised in Figure 2 that shows a schematic organisation of the locus compared to strains previously described by Hathaway *et al*. (Hathaway *et al.*, 2004). All strains had *aliB*-like ORF1, *aliB*-like ORF2, and *capN*-like regions; eight had the *doc*-like region between *capN*-like and *aliA*. Based on the classification previously proposed by Park *et al*. (Park *et al.*, 2012), the strains were therefore classified as belonging to *cps* type NCC2a (eight isolates containing the *doc*-like region) or NCC2b (the remaining five isolates). Of the eight strains belonging to *cps* type NCC2a, two had an insertion of a *tnp* region of ~1.7kb between *dexB* and *aliB*-like ORF1 previously described (Park *et al.*, 2012; Salter *et al.*, 2012).

Candidate core genome. To determine if the genome content of NT strains is comparable to that of encapsulated strains, 34 NT representing the diversity of profiles identified by PFGE, MLST, and characterisation of the capsular region, were

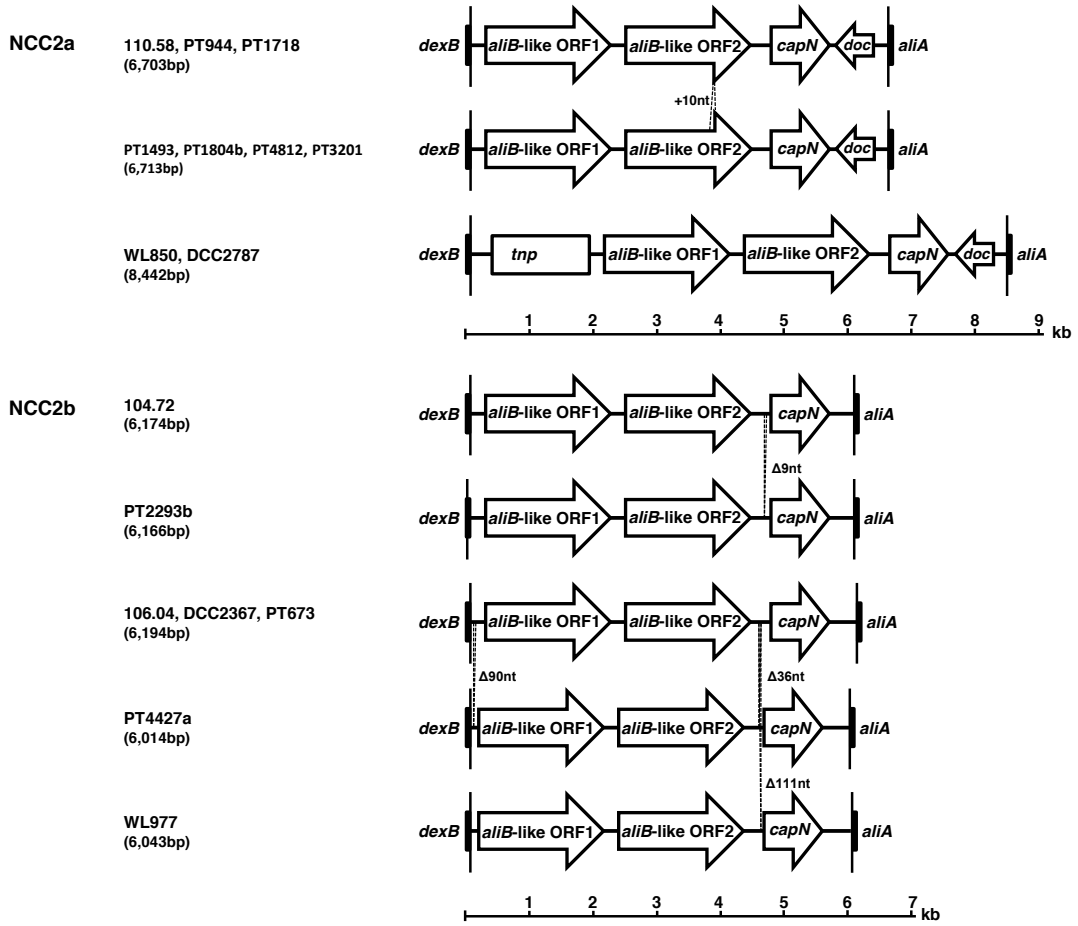


Figure 2. Schematic representation of the capsular region of NT strains.
NCC2a and NCC2b refer to a classification of *cps* types proposed by Park *et al.* (Park *et al.*, 2012). Published sequences of strains 110.58 [GenBank:AY653211.1], 104.72 [GenBank:AY653210.1], and 106.44 [GenBank:AY653209.1] are shown for comparative purposes (Hathaway *et al.*, 2004). *capN* and *doc* indicate *capN*-like and *doc*-like regions, respectively.

characterised by CGH using an array that covers the genome of nine encapsulated pneumococcal strains and R6 (a non-encapsulated derivative of D39) (Additional file 3). From the 3,052 genes present in the array, 1,666 (54.6%) were present in all NT tested, 839 (27.5%) were present in some, and 547 (17.9%) were absent in all (Additional file 4). In an independent analysis, conducted in the framework of an ongoing study, 180 encapsulated strains were analysed by CGH. These strains were representative of 20 serotypes and included all strains in the array (except R6). Results from this analysis were used for comparison. In this collection, 1,654 genes

(54.2%) were present in all strains, the same proportion found for the NT isolates. Of these 1,654 genes, 1,499 (90.6%) were also present in all NT isolates (Additional file 4). Among the remaining 155 genes, 149 were present in some (but not all) NT and only 6 were absent in all. The proportion of these 155 genes present in the NT strains ranged between 80.0% and 58.7% (Additional file 5). The 149 genes with variable presence among NT strains could be grouped into the following functions: 22.8% cellular metabolism, 16.1% transporters, 8.7% DNA metabolism, 7.4% phages and mobile elements, 2.0% surface proteins, 2.0% signalling and communication, and 41.0% were annotated as hypothetical proteins. The six genes absent in all NT were SP_0346 (annotated as capsular polysaccharide biosynthesis protein Cps4A), SP_0368 (cell wall surface anchor family protein), SP_1153 (hypothetical protein), SP_2157 (alcohol dehydrogenase, iron-containing), SP_2158 (L-fucose isomerase), and SP_2168 (fucose operon repressor, putative).

Furthermore, NT isolates contained between 2,049 and 2,120 genes detected by CGH with an average of 2,095 genes, while the 180 encapsulated strains had between 2,119 and 2,306 genes with an average of 2,235. Based on these experiments, although the size of "core" genomes of NT versus encapsulated strains was comparable, NT strains characterised in this study had 6.3% less genes detected by CGH than encapsulated strains.

Accessory regions (ARs). To further analyse the genome content of NT strains, the presence of previously identified ARs was investigated (Figure 3) (Blomberg *et al.*, 2009). Of the 41 ARs described to date, 17 were present or partially present in all NT strains analysed (ARs 3, 6, 9, 13-15, 18, 20-22, 31-33, 35, 37-39) and 7 were absent in

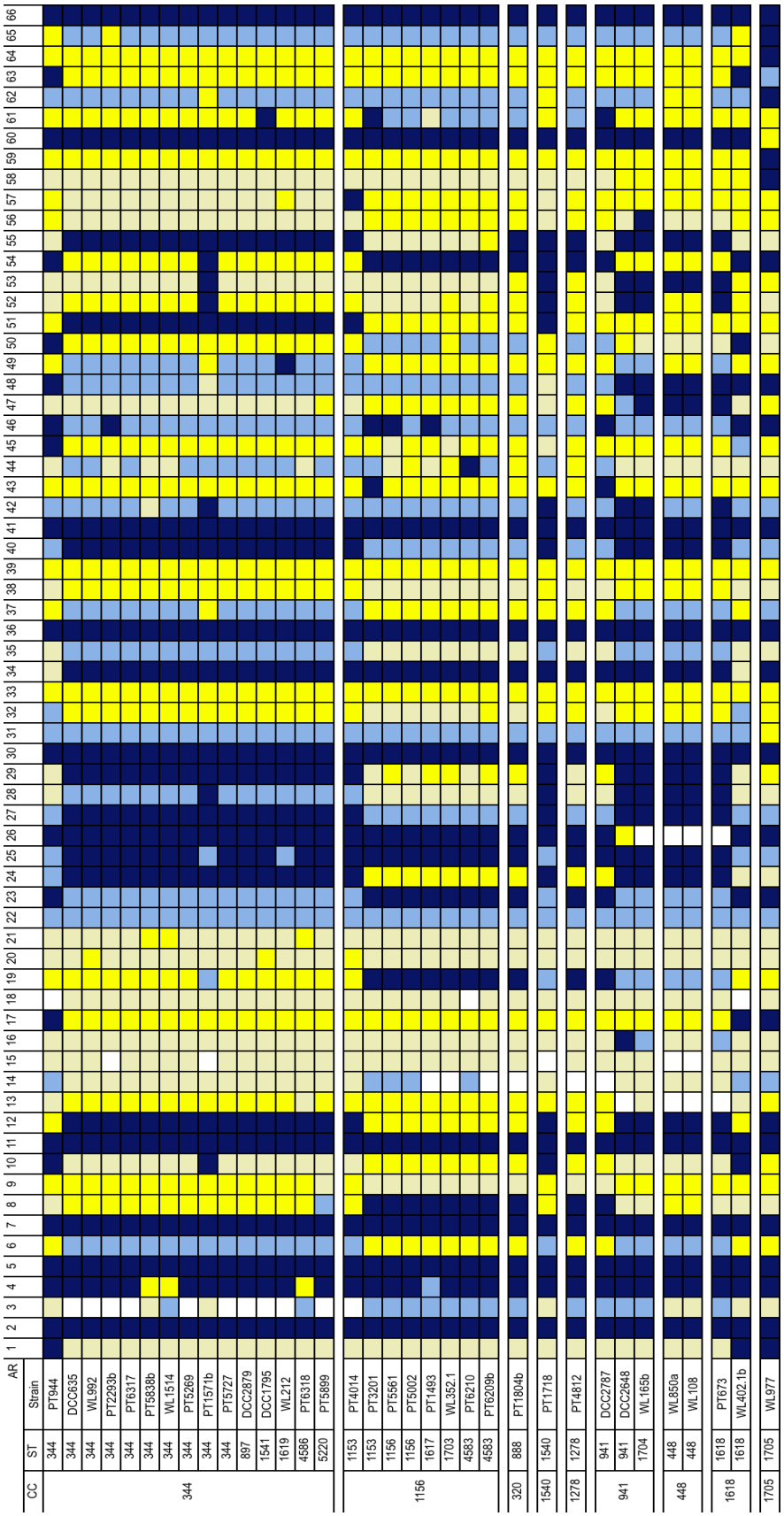


Figure 3. Distribution of accessory regions (ARs) among NT strains.
Dark yellow – AR is present; light yellow – more than 50% of the genes in the AR are present; white – 50% of the genes in the AR are present; light blue – more than 50% of the genes in the AR are absent; dark blue – AR is absent.

all (ARs 2, 5, 7, 11, 30, 36, and 41). Furthermore, 8 ARs were present, or at least partially present, in most strains (ARs 1, 8, 10, 16, 17, 19, 23, and 28) and 9 ARs were absent, or mostly absent, in most strains (ARs 4, 12, 24-27, 29, 34, and 40).

Twenty-five new ARs (named ARs 42 to 66), totalling 134 genes, were identified in this study. Their predicted functions are described in Table 2 and include ABC transporters, type II restriction-modification system, phosphotransferase system and proteins involved in metabolism, cell envelope, transport, and transcription regulation. These 25 ARs were dispersed around the TIGR4 genome (Figure 4). Of these, 22 ARs were present, or at least partially present, in most strains (ARs 42-54, 56-59, and 61-65), 2 ARs were absent, or mostly absent, in most strains (ARs 55 and 60), and AR66 (encoding for hypothetical proteins) was absent in all.

Altogether, when looking for ARs absent in all NT, these were found to encode for capsular genes (AR7), type-I pili (AR11), sucrose ABC transporter (AR36), fucose metabolism (AR41), a putative bacteriocin (AR2), and several hypothetical proteins (ARs 5, 30, and 66).

Virulence factors. A total of 496 of the genes present on the array were identified as virulence factors of pneumococcus based on published data (annotated in Additional file 4) (Polissi *et al.*, 1998; Lau *et al.*, 2001; Hava and Camilli, 2002; Garbom *et al.*, 2004; Orihuela *et al.*, 2004; LeMessurier *et al.*, 2006; Obert *et al.*, 2006; Embry *et al.*, 2007; Kadioglu *et al.*, 2008; Molzen *et al.*, 2011; Williams *et al.*, 2012). Of these, 363 (73.2%) were present in all NT strains and 36 (7.3%) were absent in all. This latter group included genes associated with capsular synthesis (TIGR4 *cpsA*, *cpsC*, *cpsD*, *cpsE*, *cpsF*, and *cpsJ*), pilus islet-1, and virulence proteins *cbpA/pspC*, *pspA*, *nanE*, *glf*,

Table 2. New accessory regions found in NT strains.

Accessory region	TIGR4 locus	Identified by STM[a]	Predicted function[b]
42	SP_0115-0117	Yes	Cell envelope
43	SP_0124-0126	No	Hypothetical
44	SP_0130-0144	Yes	ABC transporter (glucose)
45	SP_0314-0330	Yes	PTS system
46	SP_0367-0369	No	Cell envelope
47	SP_0391-0393	No	Cell envelope
48	SP_0569-0571	Yes	Type II RM system
49	SP_0595-0597	Yes	Hypothetical
50	SP_0627-0629	No	Hypothetical
51	SP_0636-0640	No	ABC transporter
52	SP_0683-0685	No	Hypothetical
53	SP_0703-0711	No	ABC transporter (aa)
54	SP_0737-0740	No	Transport & transcription regulation
55	SP_1030-1040	Yes	ABC transporter (iron)
56	SP_1042-1045	Yes	Metabolic
57	SP_1119-1125	Yes	Metabolic (glycogen)
58	SP_1160-1165	No	Metabolic (acetoin)
59	SP_1209-1211	No	Hypothetical
60	SP_1656-1658	No	Hypothetical
61	SP_1677-1679	No	Hypothetical
62	SP_1849-1851	No	Type II RM system
63	SP_1855-1859	Yes	Transport & transcription regulation
64	SP_1869-1872	Yes	ABC transporter (iron)
65	SP_2147-2154	No	Metabolic (arginine)
66	SP_2178-2183	Yes	Hypothetical

a – gene(s) within region(s) identified by signature-tagged mutagenesis as required for invasive disease Hava and Camilli, 2002; b – ATP-binding cassette (ABC); phosphotransferase (PTS); restriction modification (RM); amino acid (aa).

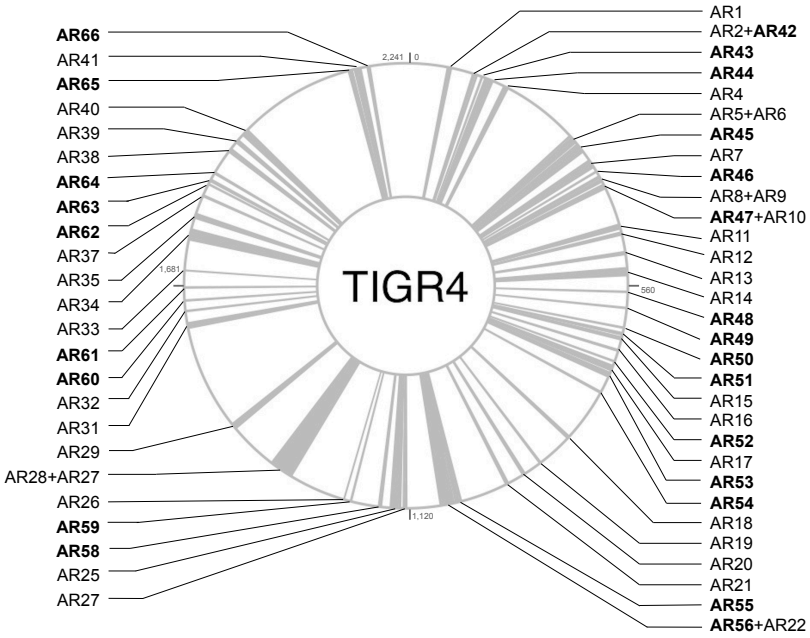


Figure 4. Distribution of 66 accessory regions (ARs) over the TIGR4 genome. Bold – new ARs identified in NT strains.

and *ntpK* among others (Additional file 4). PCR analysis of pilus-1 and -2 confirmed the absence of these loci in all 52 strains.

Regarding competence-associated genes (n=22), all were present in all strains, including the recently described *comG* operon (SP_1808 and SP_2047-53), encoding for a type-IV transformation pilus (Table 3) (Laurenceau *et al.*, 2013). In addition, *comC* and *comD* alleles were determined by PCR for the 52 NT strains included in this study and a clear distinction between CCs could be observed for *comCD*: CCs 344, 448, and 941 encoded CSP2 and ComD2; CCs 320, 1156, 1278, 1618, and 1705 had CSP1 and ComD1; and CC1540 had CSP1 and ComD4 (Iannelli *et al.*, 2005).

Nine choline binding proteins have been implicated in virulence, and all were present on the array (Polissi *et al.*, 1998; Gosink *et al.*, 2000; Hava and Camilli, 2002; Glover *et al.*, 2008). Of these, *cbpD*, *cbpE/pce*, *lytA*, *lytB*, and *lytC* were present in all strains, with *cbpA/pspC* and *pspA* being absent in all strains. Variation between CCs was found for *cbpF, cbpG* and *pcpA* (Table 3).

In addition, 12 genes implicated in colonisation were present on the array. Of these, *pavA, eno, pyrR*, *strH*, *trpG, rr01*, and *SPY2053* were present in all NT, while *rlrA* was absent in all strains. Clonal variation was found for genes *hyl*, *nanA*, *bgaA,* and *phoU* (Table 3).

Table 3. Virulence factors determined by CGH for NT clonal complexes

Gene name and/or annotation	CC344 (n=15)	CC1156 (n=8)	CC320 (n=1)	sing1540 (n=1)	sing1278 (n=1)	CC941 (n=3)	CC448 (n=2)	CC1618 (n=2)	sing1705 (n=1)
Competence proteins									
***comA*; competence factor transporting ATP-binding/permease protein ComA**	1	1	1	1	1	1	1	1	1
***comB*; competence factor transport protein ComB**	1	1	1	1	1	1	1	1	1
***comD*; putative sensor histidine kinase ComD**	1	1	1	1	1	1	1	1	1
***comE*; response regulator ComE**	1	1	1	1	1	1	1	1	1
***comX1*; transcriptional regulator ComX1**	1	1	1	1	1	1	1	1	1
competence damage-inducible protein A	1	1	1	1	1	1	1	1	1
***coiA*; competence protein CoiA**	1	1	1	1	1	1	1	1	1
competence protein ComF, putative	1	1	1	1	1	1	1	1	1
***celA*; competence protein CelA**	1	1	1	1	1	1	1	1	1
***celB*; competence protein CelB**	1	1	1	1	1	1	1	1	1
***ccs1*; competence-induced protein Ccs1**	1	1	1	1	1	1	1	1	1
***ccs4*; competence-induced protein Ccs4**	1	1	1	1	1	1	1	1	1
***ccs16*; competence-induced protein Ccs16**	1	1	1	1	1	1	1	1	1
***cspC*-related protein, authentic point mutation**	1	1	1	1	1	1	1	1	1
***pilD*; type IV prepilin peptidase, putative**	1	1	1	1	1	1	1	1	1
***comGA/ cglA*; competence protein CglA**	1	1	1	1	1	1	1	1	1
***comGB/ cglB*; competence protein CglB**	1	1	1	1	1	1	1	1	1
***comGC/ cglC*; competence protein CglC**	1	1	1	1	1	1	1	1	1
***comGD/ cglD*; competence protein CglD**	1	1	1	1	1	1	1	1	1
comGE	1	1	1	1	1	1	1	1	1
comGF	1	1	1	1	1	1	1	1	1
comGG	1	1	1	1	1	1	1	1	1
Choline-binding proteins									
***cbpA*/*pspC*; choline binding protein A**	0	0	0	0	0	0	0	0	0
***cbpD*; choline binding protein D**	1	1	1	1	1	1	1	1	1
***cbpE*/*pce*; choline binding protein E**	1	1	1	1	1	1	1	1	1
***cbpF*; choline binding protein F**	0.1	0.9	1	0	1	0.3	0	0	1
***cbpG*; choline binding protein G**	0.9	1	1	1	1	1	1	1	1
***lytA*; autolysin**	1	1	1	1	1	1	1	1	1
***lytB*; endo-beta-N-acetylglucosaminidase**	1	1	1	1	1	1	1	1	1
***lytC*; beta-N-acetylhexosaminidase**	1	1	1	1	1	1	1	1	1
***pspA*; pneumococcal surface protein A**	0	0	0	0	0	0	0	0	0
***pcpA*; choline binding protein PcpA**	0.9	0.1	0	0	0	0	0	0	0

Colonisation-associated proteins									
***hyl*; hyaluronidase**	0.9	1	1	1	1	1	1	0.5	1
***nanA*; neuraminidase A/siliase A precursor**	0.1	0	0	1	0	0	0	0.5	1
***pavA*; adherence and virulence protein A**	1	1	1	1	1	1	1	1	1
***rlrA*; transcriptional regulator, putative**	0	0	0	0	0	0	0	0	0
***bgaA*; beta-galactosidase**	0.1	0.9	1	0	1	0.3	0	0	0
***eno*; phosphopyruvate hydratase**	1	1	1	1	1	1	1	1	1
***pyrR*; bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase**	1	1	1	1	1	1	1	1	1
***strH*; beta-N-acetylhexosaminidase**	1	1	1	1	1	1	1	1	1
***trpG*; anthranilate synthase component II**	1	1	1	1	1	1	1	1	1
***phoU*; phosphate transport system regulatory protein PhoU, putative**	0.1	0.9	1	1	1	0.3	0	0.5	0
***rr01*; DNA-binding response regulator**	1	1	1	1	1	1	1	1	1
transcriptional regulator SPY2053	1	1	1	1	1	1	1	1	1
Other major virulence factors									
***ply*; pneumolysin**	1	1	1	1	1	1	1	1	1
***psaA*; manganese ABC transporter, manganese-binding adhesion liprotein**	1	1	1	1	1	1	1	1	1
***htrA*; serine protease**	1	1	1	1	1	1	1	1	1
***IgA*; immunoglobulin A1 protease**	1	1	1	1	1	1	1	1	1
***spxB*; pyruvate oxidase**	1	1	1	1	1	1	1	1	1
***piaA*; iron-compound ABC transporter, iron compound-binding protein**	0.1	0.9	0	0	0	0.3	0	0.5	1
***piaB*; iron-compound ABC transporter, permease protein**	0.1	0.9	0	0	0	0.3	0	0.5	1
***piaC*; iron-compound ABC transporter, permease protein**	0.1	0.9	0	0	0	0.3	0	0.5	1
***piaD*; iron-compound ABC transporter, ATP-binding protein**	0.1	0.9	0	0	0	0.3	0	0.5	1
***piuA*; iron-compound ABC transporter, iron-compound-binding protein**	1	1	1	1	1	1	1	1	0
***piuB*; iron-compound ABC transporter, permease protein**	1	1	1	1	1	1	1	1	0
***piuC*; iron-compound ABC transporter, permease protein**	1	1	1	1	1	1	1	1	0
***piuD*; iron-compound ABC transporter, ATP-binding protein**	1	1	1	1	1	1	1	1	0
***zmpB*; zinc metalloprotease**	0	0.8	0	0	0	0.3	0	0	1

CC – clonal complex; sing – singleton; numbers between 0 and 1 indicate the relative proportion of strains containing the gene.

Among other major virulence factors, *ply*, *psaA, htrA, IgA*, and *spxB* were present in all strains with variations between clones found for the operons *piuA-D* and *piaA-D* and *zmpB*.

Further details on the variable presence of virulence genes can be found in Additional file 4.

Intraclonal variation. Comparison of SmaI-PFGE patterns of NT strains resulted in an unexpected high diversity of profiles for strains belonging to the same ST (Figure 5) (Sá-Leão *et al.*, 2006). Likewise, there were also strains with similar PFGE profiles belonging to different STs. This lack of concordance was puzzling, as previous studies have found a good general agreement with PFGE and MLST for encapsulated pneumococci (Elberse *et al.*, 2011). To investigate possible genomic variations that could account for the lack of concordance found between PFGE and MLST results, CGH results were compared for strains belonging to the same CC. For any given CC, all strains analysed shared at least 72% of the genes detected in the NT pool (Figure 6).

When we looked at intraclonal diversity, within each CC, variation between strains was mostly due to only a few (if any) genes. Still, exceptions were found: strains PT944 of CC344, PT4014 of CC1156, and DCC2787 of CC941 had 162, 144, and 244 genes, respectively, uniquely present in their genomes compared to other strains of the same CC. Also, the two strains of CC1618 were found to differ from each other in more than 400 genes.

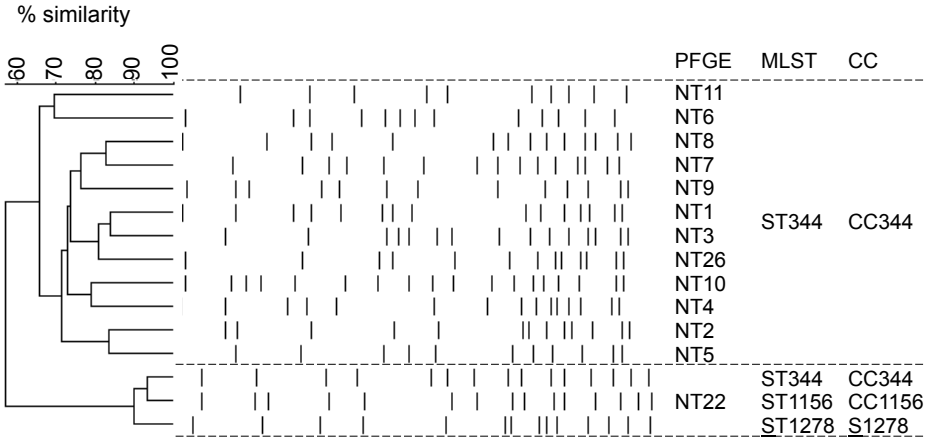


Figure 5. Comparison of PFGE patterns found for clonal complex (CC) 344, CC941, CC448, and CC1156.
Dendrogram generated by UPGMA and Dice similarity with an optimisation of 1% and a tolerance of 1.5%. CC – clonal complex; S – singleton.

When looking for the functions of genes uniquely present in one strain of a given CC, most were found to encode for hypothetical proteins (51.3%). Other genes had the following functions: transport and secretion (13.4%), cell metabolism (9.9%), phages and mobile elements (9.5%), DNA metabolism (7.8%), cell wall, cell membrane, and cell division (3.8%), signalling and communication (2.7%), and stress (1.5%). Furthermore, only 10.2% of this latter group of genes have been described as virulence genes. Not surprisingly, close to half of these genes were found in ARs (44.4%).

To investigate if the high variability of PFGE types found could be due to the presence of prophages, as previously reported (Severina *et al.*, 1999), or the presence of other mobile elements, we evaluated their distribution among NT strains (Figure 7). In some cases, e.g. NT1, NT2, and NT6 of ST344 or NT22 and NT24 of ST1153, the content of mobile elements was indeed distinct between strains, which might explain the variability found. However, in other cases, such as NT2, NT3, NT5, NT8, and NT11 of ST344 and NT15 and NT16 of ST448, the strains shared the same mobile elements. On

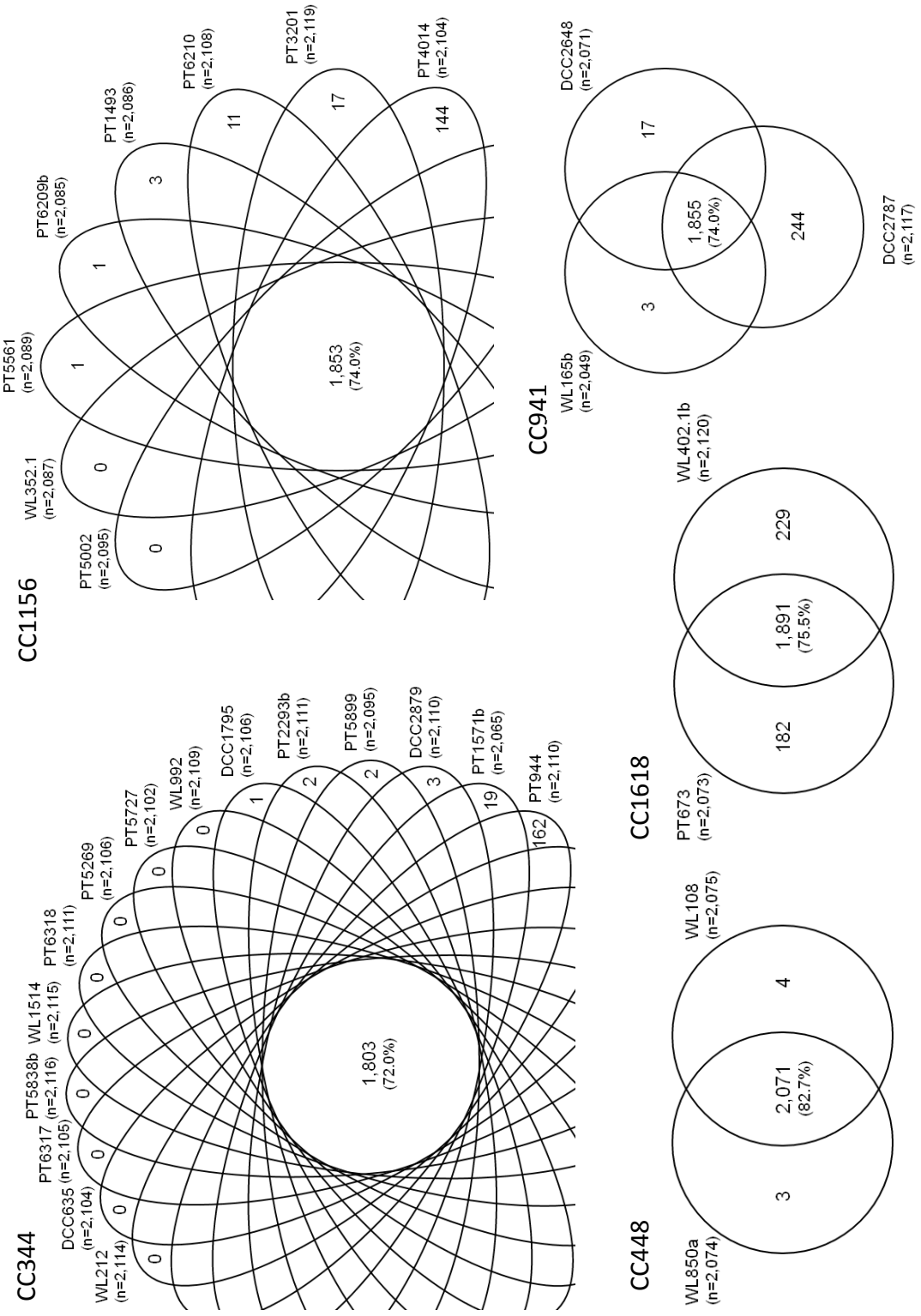


Figure 6. Intraclonal diversity of NT strains.

CC – clonal complex; numbers in the centre represent the number of genes shared by all strains of a given CC/singleton and the percentage in relation to the total number of genes detected for NT; other numbers represent the number of genes found exclusively for a given strain in comparison with strains from the same CC.

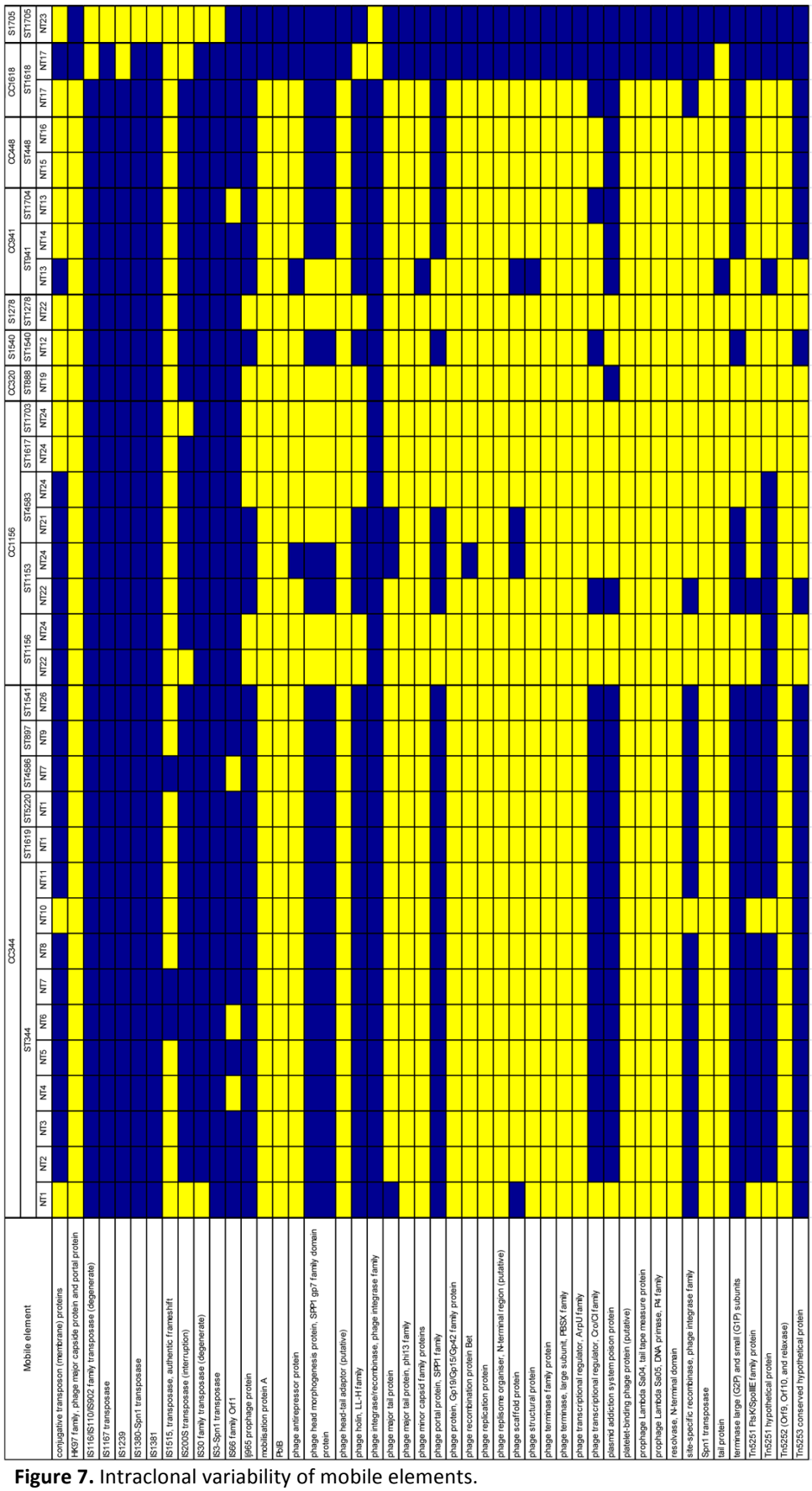


Figure 7. Intraclonal variability of mobile elements.

NT1 to NT24 refer to PFGE patterns. Yellow – present; blue – absent.

the other hand, examples of strains belonging to the same PFGE type and ST but with different mobile elements' profiles were also observed (e.g. NT17 of ST448). To complement this analysis, the presence of prophages was also determined by *lytA* hybridisation (Additional file 6). In ST344, the six PFGE types tested exhibited three *lytA* hybridisation patterns, whereas the two ST448 PFGE types tested showed the same *lytA* hybridisation pattern. According to these results, the high variability of PFGE types observed within STs could not be entirely explained by the presence of prophages or other mobile elements.

Discussion

In this study we aimed to characterise the genomic content of a collection of NT strains representative of the carriage lineages circulating in Portugal in a period of 11 years (1997-2007). Strains were analysed by CGH against a panel of 10 pneumococcal strains and their capsular region was sequenced. According to their capsular regions, strains in this study could be classified as NCC2, as they all contained *aliB*-like genes (Park *et al.*, 2012). Strains with similar capsular regions have also been identified in carriage and disease isolates circulating in Switzerland, the Netherlands, UK, USA, Brazil, South Korea, Thailand, and the Gambia (Hathaway *et al.*, 2004; Park *et al.*, 2012; Salter *et al.*, 2012; Park *et al.*, 2014).

In our collection we did not find isolates of *cps* type NCC1 (containing the *pspK*/*nspA* gene) and we did not include NT strains derived from encapsulated lineages that had

alterations in the capsular operon leading to absence of capsular production (Group I NT).

Of interest, a recent study by Park *et al*. aimed to characterise invasive NT strains from the USA. The authors reported that these strains are rare, accounting for less than 1% of the invasive pneumococcal disease cases, and most are of Group I NT, with only a few cases caused by NCC2 NT. Nonetheless, it has been clearly demonstrated that NCC2 NT are capable of causing invasive disease and therefore should not be disregarded (Hathaway *et al.*, 2004; Park *et al.*, 2014).

In relation to core genome, 54.6% of the genes represented on the array were found in all NT strains, the same proportion found for a collection of 180 encapsulated strains used for comparison (54.2%). However, the average number of total genes detected in the NT strains (2,095) was 6% less than the corresponding value found for encapsulated strains. Still, this result should be interpreted with caution as, by using a CHG approach, NT genes were probably missed to an unknown extent.

Twenty-five new ARs, dispersed around the TIGR4 genome, were identified in this study. Of the 66 ARs identified to date, only seven were absent in all NT and encoded for genes associated with sugar metabolism, capsular synthesis, type-I pilus, and hypothetical proteins (Blomberg *et al.*, 2009). Also, more than 90% of the virulence factors identified in pneumococcus were found in NT. The most relevant virulence factors absent from all NT were the capsular genes and type-I pilus (referred to above), type-II pilus, choline-binding protein A (*cbpA*/*pspC*), and pneumococcal surface protein A (*pspA*) (Kadioglu *et al.*, 2008). Also absent in the majority of NT was the major iron ABC transport system *piaA-D*. However, *piuA-D*, a second iron ABC transport system,

was present in the majority of NT. Mutations in these systems have been shown to result in mild (*piuA-D*) to moderate (*piaA-D*) reduction in virulence (Brown *et al.*, 2002). Together with the lack of capsule and other important virulence genes, the absence of these genes in NT should contribute to a lower propensity of NT to cause disease.

As expected, all strains had all competence genes, including the newly described transformation pilus (Laurenceau *et al.*, 2013; Balaban *et al.*, 2014; Chewapreecha *et al.*, 2014). According to the type of competence stimulating peptide (CSP, encoded by *comC*) secreted by pneumococcal strains, strains can be divided in pherotypes. The dominant pneumococcal pherotypes are CSP1 and CSP2, respectively found in 60-75% and 25-40% of carriage or clinical isolates (Carrolo *et al.*, 2009; Valente *et al.*, 2012). In NT, the dominant pherotype was CSP2 (65% of the strains), with the remaining strains belonging to pherotype CSP1. In our study, pherotype was a clonal property, with all strains within a CC belonging to the same pherotype. The same association was previously observed in encapsulated pneumococcus (Vestrheim *et al.*, 2011). These results further support that NT are *bona-fide* pneumococci, in contrast with atypical strains of ambiguous speciation, where multiple ComC alleles can be found (Simões *et al.*, 2010).

To explore the reasons underlying the observation that NT had highly variable PFGE profiles in contrast to relatively conserved STs, we assessed whether the presence of prophages or other mobile elements could account for these observations. Although that seemed to be the case in some strains, the presence of these mobile elements could not entirely explain the variability found in NT isolates, at least with the

approaches that were used. A more detailed characterisation of phage presence, such as the prophage typing system proposed by Romero *et al.*, could have provided additional information but was beyond the purpose of this study (Romero *et al.*, 2009a; Romero *et al.*, 2009b).

Our study has a major limitation. Information obtained by CGH is restricted to what is present in the array and therefore limited by nature. Still, interesting information regarding variability and presence/absence of pneumococcal genes implicated in virulence was obtained, providing further hypothesis related to the low disease capacity of these strains. Our study has also some strengths. The thorough characterisation of a representative collection of NT circulating in Portugal for over a decade provided insight on the most frequent features of the lineages in circulation and definitely supported the inclusion of these strains as part of the pneumococcal population.

Conclusions

NT circulating in Portugal are a homogeneous group belonging to *cps* type NCC2. Our observations support that this group are *bona-fide* pneumococcal isolates that do not express the capsule but are otherwise essentially similar to encapsulated pneumococci, having a comparable core genome and most virulence factors. Given that NT are not targeted by current pneumococcal vaccines and that they are highly transformable, we recommend that these isolates are routinely identified and reported in surveillance studies monitoring pneumococcal serotype evolution.

Acknowledgments

This work was funded by Fundação para a Ciência e a Tecnologia, Portugal, through grants PTDC/BIA-MIC/64010/2006 and PTDC/BIA-BEC/098289/2008 awarded to RSL, SFRH/BD/70147/2010 awarded to DAT, and Pest-OE/EQB/LA0004/2011 awarded to Laboratório Associado de Oeiras. The authors thank Marc J. Eleveld for technical assistance in the initial microarray experiments and Aldert Zomer for bioinformatical assistance in the array analysis. The funders had no role in the design of the study, collection, analysis, and interpretation of data, writing of the manuscript or in the decision to submit the manuscript for publication.

References

Balaban, M., P. Battig, S. Muschiol, S.M. Tirier, F. Wartha, S. Normark and B. Henriques-Normark. (2014). Secretion of a pneumococcal type II secretion system pilus correlates with DNA uptake during transformation. *Proc Natl Acad Sci USA* **111**, E758-65.

Baltz, R.H., F.H. Norris, P. Matsushima, B.S. DeHoff, P. Rockey, G. Porter, S. Burgett, R. Peery, J. Hoskins, L. Braverman, I. Jenkins, P. Solenberg, M. Young, M.A. McHenney, P.L. Skatrud and P.R. Rosteck, Jr. (1998). DNA sequence sampling of the *Streptococcus pneumoniae* genome to identify novel targets for antibiotic development. *Microb Drug Resist* **4**, 1-9.

Bentley, S.D., D.M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M.A. Quail, G. Samuel, I.C. Skovsted, M.S. Kaltoft, B. Barrell, P.R. Reeves, J. Parkhill and B.G. Spratt. (2006). Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2**, e31.

Blomberg, C., J. Dagerhamn, S. Dahlberg, S. Browall, J. Fernebro, B. Albiger, E. Morfeldt, S. Normark and B. Henriques-Normark. (2009). Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. *J Infect Dis* **199**, 1032-42.

Browall, S., M. Norman, J. Tangrot, I. Galanis, K. Sjostrom, J. Dagerhamn, C. Hellberg, A. Pathak, T. Spadafina, A. Sandgren, P. Battig, O. Franzen, B. Andersson, A. Ortqvist, S. Normark and B. Henriques-Normark. (2014). Intraclonal variations among *Streptococcus pneumoniae* isolates influence the likelihood of invasive disease in children. *J Infect Dis* **209**, 377-88.

Brown, J.S., S.M. Gilliland, J. Ruiz-Albert and D.W. Holden. (2002). Characterization of *pit*, a *Streptococcus pneumoniae* iron uptake ABC transporter. *Infect Immun* **70**, 4389-98.

Carrolo, M., F.R. Pinto, J. Melo-Cristino and M. Ramirez. (2009). Pherotypes are driving genetic differentiation within *Streptococcus pneumoniae*. *BMC Microbiol* **9**, 191.

Chewapreecha, C., S.R. Harris, N.J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D.M. Aanensen, A.E. Mather, A.J. Page, S.J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner and S.D. Bentley. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46**, 305-9.

Domenech, M., E. Garcia and M. Moscoso. (2009). Versatility of the capsular genes during biofilm formation by *Streptococcus pneumoniae*. *Environ Microbiol* **11**, 2542-55.

Elberse, K.E., S. Nunes, R. Sá-Leão, H.G. van der Heide and L.M. Schouls. (2011). Multiple-locus variable number tandem repeat analysis for *Streptococcus pneumoniae*: comparison with PFGE and MLST. *PLoS One* **6**, e19668.

Embry, A., E. Hinojosa and C.J. Orihuela. (2007). Regions of Diversity 8, 9 and 13 contribute to *Streptococcus pneumoniae* virulence. *BMC Microbiol* **7**, 80.

Garbom, S., A. Forsberg, H. Wolf-Watz and B.M. Kihlberg. (2004). Identification of novel virulence-associated genes via genome analysis of hypothetical genes. *Infect Immun* **72**, 1333-40.

Glover, D.T., S.K. Hollingshead and D.E. Briles. (2008). *Streptococcus pneumoniae* surface protein PcpA elicits protection against lung infection and fatal sepsis. *Infect Immun* **76**, 2767-76.

Gosink, K.K., E.R. Mann, C. Guglielmo, E.I. Tuomanen and H.R. Masure. (2000). Role of novel choline binding proteins in virulence of *Streptococcus pneumoniae*. *Infect Immun* **68**, 5690-5.

Hanage, W.P., T. Kaijalainen, E. Herva, A. Saukkoriipi, R. Syrjanen and B.G. Spratt. (2005). Using multilocus sequence data to define the pneumococcus. *J Bacteriol* **187**, 6223-30.

Hathaway, L.J., P. Stutzmann Meier, P. Battig, S. Aebi and K. Muhlemann. (2004). A homologue of *aliB* is found in the capsule region of nonencapsulated *Streptococcus pneumoniae*. *J Bacteriol* **186**, 3721-9.

Hava, D.L. and A. Camilli. (2002). Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Mol Microbiol* **45**, 1389-406.

Hiller, N.L., B. Janto, J.S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N.E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S.D. Bentley, J.C. Post, G.D. Ehrlich and F.Z. Hu. (2007). Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol* **189**, 8186-95.

Hyams, C., S. Opel, W. Hanage, J. Yuste, K. Bax, B. Henriques-Normark, B.G. Spratt and J.S. Brown. (2011). Effects of *Streptococcus pneumoniae* strain background on complement resistance. *PLoS One* **6**, e24581.

Iannelli, F., M.R. Oggioni and G. Pozzi. (2005). Sensor domain of histidine kinase ComD confers competence pherotype specificity in *Streptococcoccus pneumoniae*. *FEMS Microbiol Lett* **252**, 321-6.

Iannelli, F., B.J. Pearce and G. Pozzi. (1999). The type 2 capsule locus of *Streptococcus pneumoniae*. *J Bacteriol* **181**, 2652-4.

Kadioglu, A., J.N. Weiser, J.C. Paton and P.W. Andrew. (2008). The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat Rev Microbiol* **6**, 288-301.

Keller, L.E., C.V. Jones, J.A. Thornton, M.E. Sanders, E. Swiatlo, M.H. Nahm, I.H. Park and L.S. McDaniel. (2013). PspK of *Streptococcus pneumoniae* increases adherence to epithelial cells and enhances nasopharyngeal colonization. *Infect Immun* **81**, 173-81.

Kellogg, J.A., D.A. Bankert, C.J. Elder, J.L. Gibbs and M.C. Smith. (2001). Identification of *Streptococcus pneumoniae* revisited. *J Clin Microbiol* **39**, 3373-5.

Kilian, M., K. Poulsen, T. Blomqvist, L.S. Havarstein, M. Bek-Thomsen, H. Tettelin and U.B. Sorensen. (2008). Evolution of *Streptococcus pneumoniae* and its close commensal relatives. *PLoS One* **3**, e2683.

Lau, G.W., S. Haataja, M. Lonetto, S.E. Kensit, A. Marra, A.P. Bryant, D. McDevitt, D.A. Morrison and D.W. Holden. (2001). A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. *Mol Microbiol* **40**, 555-71.

Laurenceau, R., G. Pehau-Arnaudet, S. Baconnais, J. Gault, C. Malosse, A. Dujeancourt, N. Campo, J. Chamot-Rooke, E. Le Cam, J.P. Claverys and R. Fronzes. (2013). A type IV pilus mediates DNA binding during natural transformation in *Streptococcus pneumoniae*. *PLoS Pathog* **9**, e1003473.

LeMessurier, K.S., A.D. Ogunniyi and J.C. Paton. (2006). Differential expression of key pneumococcal virulence genes *in vivo*. *Microbiology* **152**, 305-11.

Magee, A.D. and J. Yother. (2001). Requirement for capsule in colonization by *Streptococcus pneumoniae*. *Infect Immun* **69**, 3755-61.

Martin, M., J.H. Turco, M.E. Zegans, R.R. Facklam, S. Sodha, J.A. Elliott, J.H. Pryor, B. Beall, D.D. Erdman, Y.Y. Baumgartner, P.A. Sanchez, J.D. Schwartzman, J. Montero, A. Schuchat and C.G. Whitney. (2003). An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. *N Engl J Med* **348**, 1112-21.

Molzen, T.E., P. Burghout, H.J. Bootsma, C.T. Brandt, C.E. van der Gaast-de Jongh, M.J. Eleveld, M.M. Verbeek, N. Frimodt-Moller, C. Ostergaard and P.W. Hermans. (2011). Genome-wide identification of *Streptococcus pneumoniae* genes essential for bacterial replication during experimental meningitis. *Infect Immun* **79**, 288-97.

Nuorti, J.P. and C.G. Whitney. (2010). Prevention of pneumococcal disease among infants and children - use of 13-valent pneumococcal conjugate vaccine and 23-valent pneumococcal polysaccharide vaccine - recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep* **59**, 1-18.

Obert, C., J. Sublett, D. Kaushal, E. Hinojosa, T. Barton, E.I. Tuomanen and C.J. Orihuela. (2006). Identification of a Candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun* **74**, 4766-77.

Oliver, M.B., M.P. van der Linden, S.A. Kuntzel, J.S. Saad and M.H. Nahm. (2013). Discovery of *Streptococcus pneumoniae* serotype 6 variants with glycosyltransferases synthesizing two differing repeating units. *J Biol Chem* **288**, 25976-85.

Orihuela, C.J., J.N. Radin, J.E. Sublett, G. Gao, D. Kaushal and E.I. Tuomanen. (2004). Microarray analysis of pneumococcal gene expression during invasive disease. *Infect Immun* **72**, 5582-96.

Park, I.H., K.A. Geno, L.K. Sherwood, M.H. Nahm and B. Beall. (2014). Population-based analysis of invasive nontypeable pneumococci reveals that most have defective capsule synthesis genes. *PLoS One* **9**, e97825.

Park, I.H., K.H. Kim, A.L. Andrade, D.E. Briles, L.S. McDaniel and M.H. Nahm. (2012). Nontypeable pneumococci can be divided into multiple *cps* types, including one type expressing the novel gene *pspK*. *MBio* **3**,

Polissi, A., A. Pontiggia, G. Feger, M. Altieri, H. Mottl, L. Ferrari and D. Simon. (1998). Large-scale identification of virulence genes from *Streptococcus pneumoniae*. *Infect Immun* **66**, 5620-9.

Rodrigues, F., S. Nunes, R. Sa-Leao, G. Goncalves, L. Lemos and H. de Lencastre. (2009). *Streptococcus pneumoniae* nasopharyngeal carriage in children attending day-care centers in the central region of Portugal, in the era of 7-valent pneumococcal conjugate vaccine. *Microb Drug Resist* **15**, 269-77.

Rolo, D., S.S. A, A. Domenech, A. Fenoll, J. Linares, H. de Lencastre, C. Ardanuy and R. Sá-Leão. (2013). Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. *PLoS One* **8**, e57047.

Romero, P., N.J. Croucher, N.L. Hiller, F.Z. Hu, G.D. Ehrlich, S.D. Bentley, E. Garcia and T.J. Mitchell. (2009a). Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. *J Bacteriol* **191**, 4854-62.

Romero, P., E. Garcia and T.J. Mitchell. (2009b). Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. *Appl Environ Microbiol* **75**, 1642-9.

Sá-Leão, R., S. Nunes, A. Brito-Avô, C.R. Alves, J.A. Carriço, J. Saldanha, J.S. Almeida, I. Santos-Sanches and H. de Lencastre. (2008). High rates of transmission of and colonization by *Streptococcus pneumonia*e and *Haemophilus influenzae* within a day care center revealed in a longitudinal study. *J Clin Microbiol* **46**, 225-34.

Sá-Leão, R., A.S. Simões, S. Nunes, N.G. Sousa, N. Frazão and H. de Lencastre. (2006). Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. *Microbiology* **152**, 367-76.

Sá-Leão, R., A. Tomasz, I.S. Sanches, S. Nunes, C.R. Alves, A.B. Avô, J. Saldanha, K.G. Kristinsson and H. de Lencastre. (2000). Genetic diversity and clonal patterns among antibiotic-susceptible and -resistant *Streptococcus pneumoniae* colonizing children: day care centers as autonomous epidemiological units. *J Clin Microbiol* **38**, 4137-44.

Salter, S.J., J. Hinds, K.A. Gould, L. Lambertsen, W.P. Hanage, M. Antonio, P. Turner, P.W. Hermans, H.J. Bootsma, K.L. O'Brien and S.D. Bentley. (2012). Variation at the capsule locus, *cps*, of mistyped and non-typable *Streptococcus pneumoniae* isolates. *Microbiology* **158**, 1560-9.

Scott, J.R., J. Hinds, K.A. Gould, E.V. Millar, R. Reid, M. Santosham, K.L. O'Brien and W.P. Hanage. (2012). Nontypeable pneumococcal isolates among navajo and white mountain apache communities: are these really a cause of invasive disease? *J Infect Dis* **206**, 73-80.

Severina, E., M. Ramirez and A. Tomasz. (1999). Prophage carriage as a molecular epidemiological marker in *Streptococcus pneumoniae*. *J Clin Microbiol* **37**, 3308-15.

Simões, A.S., L. Pereira, S. Nunes, A. Brito-Avo, H. de Lencastre and R. Sá-Leão. (2011a). Clonal evolution leading to maintenance of antibiotic resistance rates among colonizing Pneumococci in the PCV7 era in Portugal. *J Clin Microbiol* **49**, 2810-7.

Simões, A.S., R. Sá-Leão, M.J. Eleveld, D.A. Tavares, J.A. Carriço, H.J. Bootsma and P.W. Hermans. (2010). Highly penicillin-resistant multidrug-resistant pneumococcus-like strains colonizing children in Oeiras, Portugal: genomic characteristics and implications for surveillance. *J Clin Microbiol* **48**, 238-46.

Simões, A.S., C. Valente, H. de Lencastre and R. Sá-Leão. (2011b). Rapid identification of noncapsulated *Streptococcus pneumoniae* in nasopharyngeal samples allowing detection of co-colonization and reevaluation of prevalence. *Diagn Microbiol Infect Dis* **71**, 208-16.

Tettelin, H. and S.K. Hollingshead, 2004. Comparative genomics of *Streptococcus pneumoniae*: intrastrain diversity and genome plasticity. In: The pneumococcus, E. I. Tuomanen, (Ed.). ASM Press, Washington, D. C.: pp: 15-29.

Tettelin, H., K.E. Nelson, I.T. Paulsen, J.A. Eisen, T.D. Read, S. Peterson, J. Heidelberg, R.T. DeBoy, D.H. Haft, R.J. Dodson, A.S. Durkin, M. Gwinn, J.F. Kolonay, W.C. Nelson, J.D. Peterson, L.A. Umayam, O. White, S.L. Salzberg, M.R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A.M. Wolf, T.R. Utterback, C.L.

Hansen, L.A. McDonald, T.V. Feldblyum, S. Angiuoli, T. Dickinson, E.K. Hickey, I.E. Holt, B.J. Loftus, F. Yang, H.O. Smith, J.C. Venter, B.A. Dougherty, D.A. Morrison, S.K. Hollingshead and C.M. Fraser. (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**, 498-506.

Valente, C., H. de Lencastre and R. Sá-Leão. (2012). Pherotypes of co-colonizing pneumococci among Portuguese children. *Microb Drug Resist* **18**, 550-4.

Vestrheim, D.F., P. Gaustad, I.S. Aaberge and D.A. Caugant. (2011). Pherotypes of pneumococcal strains co-existing in healthy children. *Infect Genet Evol* **11**, 1703-8.

Weiser, J.N. and M. Kapoor. (1999). Effect of intrastrain variation in the amount of capsular polysaccharide on genetic transformation of *Streptococcus pneumoniae*: implications for virulence studies of encapsulated strains. *Infect Immun* **67**, 3690-2.

Whalan, R.H., S.G. Funnell, L.D. Bowler, M.J. Hudson, A. Robinson and C.G. Dowson. (2006). Distribution and genetic diversity of the ABC transporter lipoproteins PiuA and PiaA within *Streptococcus pneumoniae* and related streptococci. *J Bacteriol* **188**, 1031-8.

Whatmore, A.M., V.A. Barcus and C.G. Dowson. (1999). Genetic diversity of the streptococcal competence (*com*) gene locus. *J Bacteriol* **181**, 3144-54.

Williams, T.M., N.J. Loman, C. Ebruke, D.M. Musher, R.A. Adegbola, M.J. Pallen, G.M. Weinstock and M. Antonio. (2012). Genome analysis of a highly virulent serotype 1 strain of *Streptococcus pneumoniae* from West Africa. *PLoS One* **7**, e26742.

Xu, Q., R. Kaur, J.R. Casey, V. Sabharwal, S. Pelton and M.E. Pichichero. (2011). Nontypeable *Streptococcus pneumoniae* as an otopathogen. *Diagn Microbiol Infect Dis* **69**, 200-4.

Zahner, D., A. Gudlavalleti and D.S. Stephens. (2010). Increase in pilus islet 2-encoded pili among *Streptococcus pneumoniae* isolates, Atlanta, Georgia, USA. *Emerg Infect Dis* **16**, 955-62.

Supporting information

NOTE: Additional files are not presented in the same order as in the original paper.

Additional file 1. Primers used to amplify the capsular region of NT strains.

Primer	Sequence	Reference
cap F	GTTKTGGCTAACTTGTCCAATG	(Kilian *et al.*, 2008)
cps 1R	TGAGCTGTATAGGCGTGGCG	this paper
cps 1F	AATCGCCACGCCTATACAGC	this paper
cps 2R	GATACTGAGCATTCGATTCC	this paper
cps 2F	AGCGGCCTTCAAAAAGTGGC	this paper
cps 3R	TACGGATGGTCAATTCTTGG	this paper
cps 3F	GACCAAGAATTGACCATCCG	this paper
cps 4R	ATCCAAGCCTGTGCTTCAGC	this paper
cps 4F	ATGCTGAAGCACAGGCTTGG	this paper
cps 5R	ATATGAGCCGTTTGCTCAGC	this paper
cps 5F	AAGATTGCTGAGCAAACGGC	this paper
cps 6R	GCAGCTAAAACACCAGCTGC	this paper
cps 6F	GCAGCTGGTGTTTTAGCTGC	this paper
cap R	CTGTCAACCAAGCTTGGGC	(Kilian *et al.*, 2008)
cps 7F	AAGTGGCTCTTAGGAGCAGG	this paper
cps 7F_2	GGTAAAATGTCAAGCGACCC	this paper
cps 7R	GAGTTGGCGCTCCGTAATCC	this paper
cps 8F	GTCTATAGTCCACAAGAGGC	this paper
cps 9F	TACCAAAGCTTATTCACAGG	this paper
cps10F	CACGCCCAGAACCTTACTGG	this paper
cps 11R	GCTTTCTTGCTCCCATTTGG	this paper
cps 12F	AGCTAATTACAAGGGTAGCC	this paper
cps 12R	AAAGGGTGGAAGGTCAGTCG	this paper
cps 13F	GTTAGAATACCGTAGTCTTCG	this paper
cps 13R	ACTGGTGTTGACGAGGTGGC	this paper
cps 14R	TTAGTTTCAATCACCGAAGC	this paper
cps 15F	GCTATCAACCATACGAGC	this paper
cps 16F	GGAAACAGCTAGTCTGTTGG	this paper
cps 16R	GCTTGCTCAAACCATTCAAGC	this paper
cps 17F	CAGAAGAGMAYYTMTGTTGGC	this paper
cps 17F_2	CAGAAGAGAACCTATGTTGGC	this paper
cps 18F	GTTAGCAAGTTCGTCTAAGG	this paper
cps 18R	TTCTGCATCTAGTAGGATGC	this paper
cps 19F	CCTCTAGCTAATTACAAGGG	this paper
cps 19R	TGATGATAAAGGGTGGAAGG	this paper
cps 20F	GTATTCTCTATAGCGGACC	this paper
cps 20R	GTGTGATTGTAAGCCTTACG	this paper

Additional file 2. Validation of the microarray.

Strain	No. ORFs in array	False negatives	Description
TIGR4	2,238	SP_1217	Hypothetical protein
		SP_1921	Hypothetical protein
R6	113	Not hybridised[a]	
D39	13	None	
BHN100	192	BHN100_0544	Glutamyl aminopeptidase
		BHN100_1337	Caax amino protease
		BHN100_2111	Transposase
		BHN100_2148	Transposase
		BHN100_2155	Transposase
CBR206	170	CBR206_1142	Tn5252
		CBR206_1354	Hypothetical protein
		CBR206_2210	Hypothetical protein
		CBR206_2248	Phage holin
LGST215	4	None	
BHN191	178	BHN191_0760	EcoBI specificity protein
		BHN191_0783	Hypothetical protein
		BHN191_2196	Cell wall surface anchor family protein
BHN418	1	None	
Sp14-BS69	92	SP14_0054	Gp18
		SP14_1091	Hypothetical protein
Sp3-BS71	51	None	
Total ORFs	3,052		

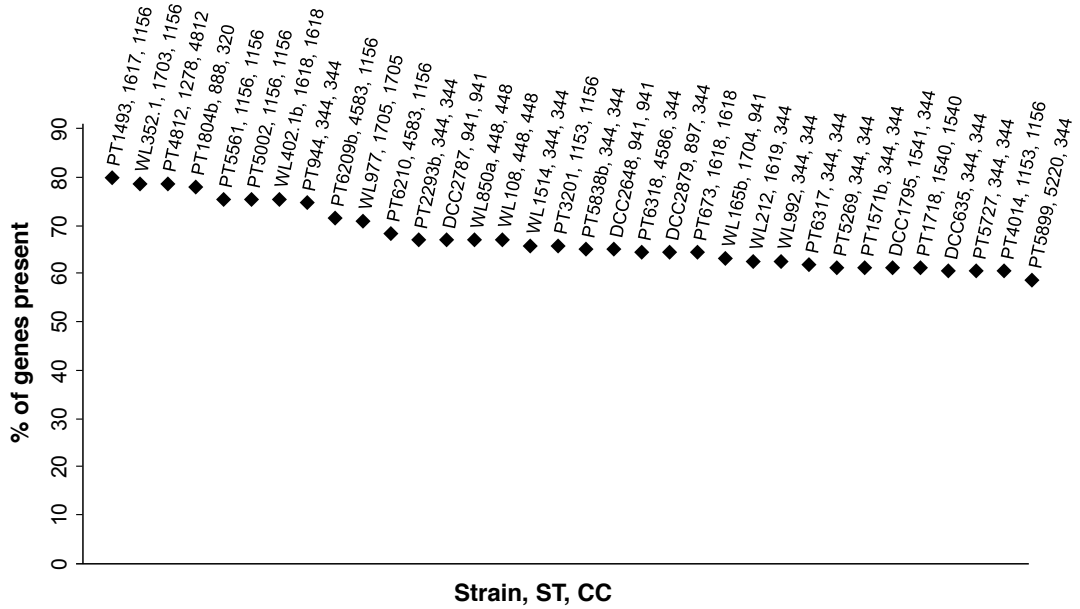
a – R6 is a derivative of D39 and was not hybridised.

Additional file 3. Strains represented in the array.

Strain	Serotype	MLST	Reference
TIGR4	4	205	(Tettelin *et al.*, 2001)
D39	2	595	(Iannelli *et al.*, 1999)
R6	-	595	(Baltz *et al.*, 1998)
CBR206	19F	179	(Rodrigues *et al.*, 2009)
LGST215	19F	179	(Sá-Leão *et al.*, 2008)
Sp3-BS71	3	180	(Hiller *et al.*, 2007)
Sp14-BS69	14	124	(Hiller *et al.*, 2007)
BHN100	19F	162	(Hyams *et al.*, 2011)
BHN191	6B	138	(Browall *et al.*, 2014)
BHN418	6B	138	(Browall *et al.*, 2014)

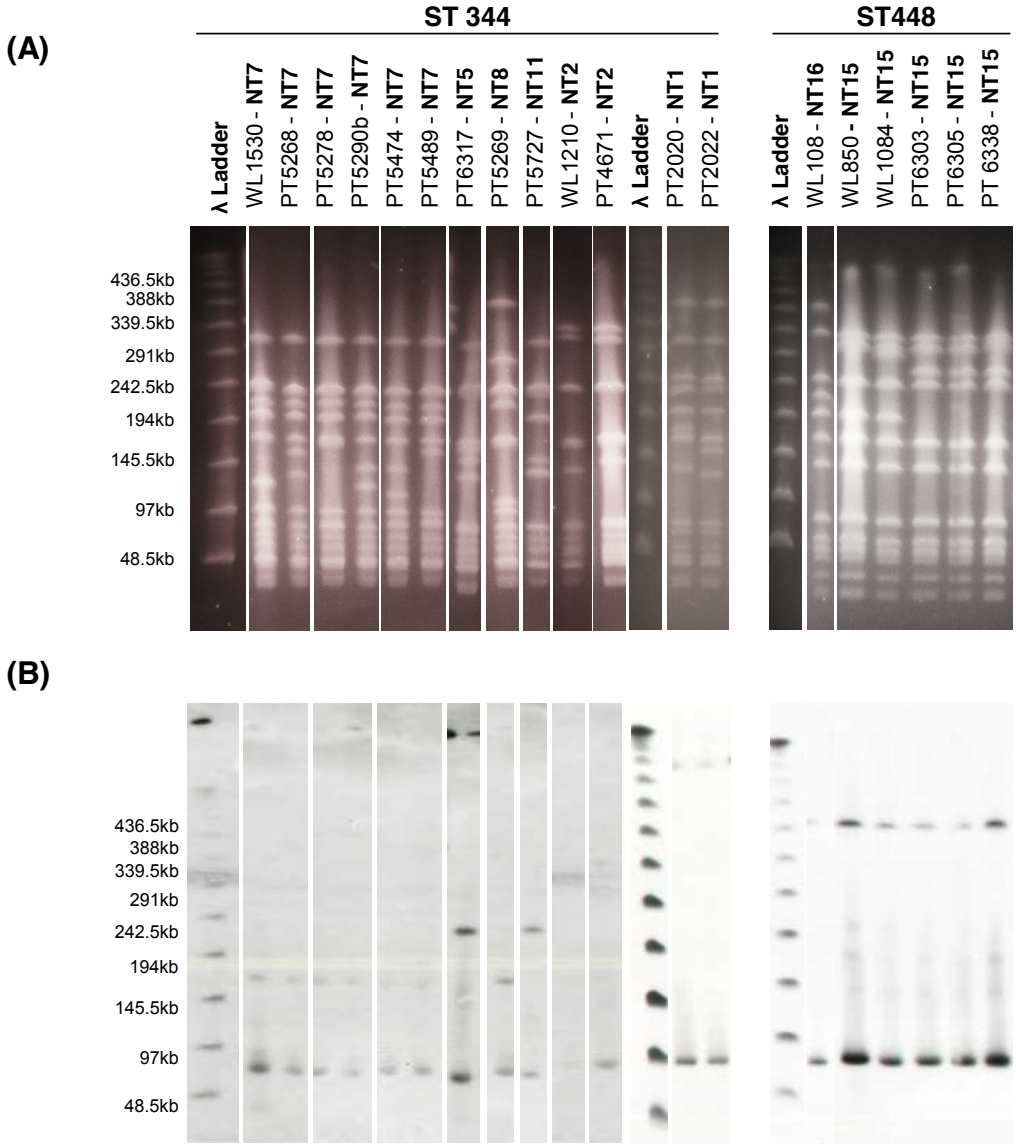
Additional file 4 (.xlsx). Core genome, virulence genes, and accessory regions.
a – annotations for: TIGR4 (SP_), D39 (SPN_), R6 (spr), CBR206 (CBR206_), LGST215 (DCCPN215_), Sp3-BS71 (SP3_), Sp14-BS69 (SP14_), BHN100 (BHN100_), BHN191 (BHN191_), and BHN418 (BHN418_); Red – genes present in all NT strains analysed; bold – new accessory regions identified in NT strains.

Please refer to the enclosed CD.



Additional file 5. Percentage of the 155 genes absent in some NT but present in a group of 180 diverse encapsulated strains (see text).

ST – multi-locus sequence type; CC – clonal complex.



Additional file 6. Detection of prophages by *lytA* hybridisation.

A – SmaI-PFGE patterns of strains representing ST344 and ST448; B – southern blotting of the PFGE gel with a probe for *lytA*.

Chapter 3

***lytA*-based identification methods can misidentify *Streptococcus pneumonia*e**

Published in: A. S. Simões*, D. A. Tavares*, D. Rolo, C. Ardanuy, H. Goossens, B. Henriques-Normark, J. Linares, H. de Lencastre, and R. Sá-Leão (2016) *Diagn Microbiol Infect Dis* **doi:10.1016/j.diagmicrobio.2016.03.18** [Epub ahead of print].

*Equal contribution.

Contributions:

D. A. Tavares was responsible for all experimental work, with the exception of the initial selection of the strains, multilocus sequence typing of the strains, and deposit of nucleotide sequences to the GenBank database, which was performed by A. S. Simões.

Summary

During surveillance studies we detected, among over 1500 presumptive pneumococci, 11 isolates displaying conflicting or novel results when characterized by widely accepted phenotypic (optochin susceptibility and bile solubility) and genotypic (*lytA*-BsaAI-RFLP and MLST) identification methods. We aimed to determine the genetic basis for the unexpected results given by *lytA*-BsaAI-RFLP and investigate the accuracy of the WHO recommended *lytA* real-time PCR assay to classify these 11 isolates. Three novel *lytA*-BsaAI-RFLP signatures were found (one in pneumococcus and two in *S. mitis*). In addition, one pneumococcus displayed the atypical *lytA*-BsaAI-RFLP signature characteristic of non-pneumococci and two *S. pseudopneumoniae* displayed the typical *lytA*-BsaAI-RFLP pattern characteristic of pneumococci. *lytA* real-time PCR misidentified these three isolates. In conclusion, identification of pneumococci by *lytA* real-time PCR, and other *lytA*-based methodologies, may lead to false results. This is of particular relevance in the increasingly frequent colonization studies relying solely on culture-independent methods.

Introduction

Streptococcus pneumoniae (pneumococcus) is a major human pathogen, causing a wide range of infections from otitis media to bacteremia and meningitis. Routine identification of pneumococcus (colony morphology on blood agar plates, susceptibility to optochin, cell wall lysis by 1% of sodium deoxycholate (bile

solubility), and assignment of a capsular type by serotyping) is not always straightforward since some isolates may give atypical results in one or more of these assays (Arbique *et al.*, 2004; Balsalobre *et al.*, 2006; Bosshard *et al.*, 2004; Nunes *et al.*, 2008; Sá-Leão *et al.*, 2006; Simões *et al.*, 2010; Whatmore *et al.*, 2000).

As a human colonizer, pneumococci co-habit the nasopharynx with several other bacterial species, including its closest relatives: *S. pseudopneumoniae*, *S. mitis*, and *S. oralis*. The exchange of genetic elements between pneumococci and its closest relatives has been described and increases the difficulties in species identification (Denapaite *et al.*, 2010; Donati *et al.*, 2010; Johnston *et al.*, 2010). Although isolates of closely related species have been implied in disease episodes, pneumococcus is the most important disease-causing species of the mitis group (formed by pneumococcus, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis*, among others) (Bochud *et al.*, 1994; Douglas *et al.*, 1993; Keith *et al.*, 2006; Rolo *et al.*, 2013). For this reason, a correct identification of pneumococcus is crucial for an accurate diagnosis and treatment. In fact, misidentification of pneumococcus could falsely increase the rates of pneumococci non-susceptible to antimicrobials since high rates of penicillin-resistant and multidrug-resistant *S. mitis* isolates have been described (Ioannidou *et al.*, 2001; Simões *et al.*, 2010; Wester *et al.*, 2002).

In recent years, several molecular methods have been proposed to differentiate pneumococcus from closely related species. The presence of the *lytA* gene – the major autolysin and a ubiquitous virulence factor – has been proposed to identify pneumococci (Messmer *et al.*, 1997). However, homologues of the *lytA* gene have been detected in strains of closely related streptococcal species (Denapaite *et al.*,

2010; Romero *et al.*, 2004; Whatmore *et al.*, 2000). A *lytA*-BsaAI-RFLP strategy to differentiate pneumococcus from closely related species based on signatures characteristic of pneumococcal (typical) *lytA* or non-pneumococcal (atypical) *lytA* has been proposed and successfully used (Llull *et al.*, 2006). Also, based on DNA sequence differences between the pneumococcal *lytA* and its homogues, real-time PCR assays for the specific identification of pneumococcus have been developed (Carvalho Mda *et al.*, 2007). Nowadays, the *lytA* real-time PCR strategy developed by the CDC is currently the WHO recommended culture-independent method to detect pneumococci (Carvalho Mda *et al.*, 2007; Satzke *et al.*, 2013).

In addition, multilocus sequence typing (MLST) and multilocus sequence analysis (MLSA) strategies have been validated as tools for reliable species identification among streptococci of the viridans group (Hanage *et al.*, 2005a; Bishop *et al.*, 2009).

During surveillance studies, we detected 11 presumptive pneumococcal isolates displaying conflicting or novel results when characterized by the combination of optochin susceptibility, bile solubility, *lytA*-BsaAI-RFLP, and MLST. In this study, we aimed to determine the genetic basis for the unexpected results given by *lytA*-BsaAI-RFLP. Also, considering the increasing and wide use of *lytA* real-time PCR for the identification of pneumococci, we also aimed to investigate the accuracy of this method in the classification of these 11 isolates.

Materials and methods

Ethics statement. In the present study a sub-set of bacterial isolates selected from different studies was characterized. All samples have been coded numerically upon collection and processed anonymously. No human subjects, human material or human data were used, thus excusing the requirement for an ethical approval. Approval for the original studies was obtained from: i) the Portuguese Ministry of Education; the study was registered and approved at the Health Care Centre of Oeiras that reports to Administração Regional de Saúde (ARS; "Regional Health Administration") of Lisboa and Vale do Tejo from the Ministry of Health (PT coded isolates); signed informed consent was obtained from parents/guardians of participating children; ii) the "Comité Ètic d'Investigació Clínica del Hospital Universitari de Bellvitge" (Spain coded isolates); and iii) research sites involved in the European project GRACE (Genomics to Combat Resistance against Antibiotics in Community-acquired LRTI in Europe) obtained ethical and competent authority approval from their local organizations. Patients who fulfilled the inclusion criteria were given written and verbal information on the study and asked for informed consent (GRA coded isolates).

Study isolates. Isolates were selected from biological samples under the framework of other studies aimed to identify pneumococci. These included surveillance carriage studies performed in Portugal (obtained between 2011-2014, n=1226 isolates), a study of lower respiratory tract infections in several European countries (obtained between 2007-2010, n=204 isolates; GRACE - Genomics to Combat Resistance against Antibiotics in Community-acquired LRTI in Europe; www.gracelrti.org), and a

collection of disease isolates with atypical properties from Spain (obtained between 1991-2009, n=132; (Rolo *et al*., 2013)). The assays performed for species identification are described below and are summarized in Fig. S1. Briefly, isolates were presumptively identified as pneumococci based on presence of a-hemolysis when grown in gentamycin blood agar plates and optochin susceptibility. If optochin resistance was observed, bile solubility was performed. Presumptive pneumococci were then serotyped by the Quellung reaction and/or by multiplex PCR. When the assignment of a serotype was not possible, a multiplex PCR designed to detect non-encapsulated pneumococci and *lytA*-BsaAI-RFLP were performed as described below (Simões *et al*., 2011).

Among the 1562 isolates mentioned above and presumptively identified as pneumococci, *lytA*-BsaAI-RFLP was performed for 247 isolates (99 from Portugal, 25 from GRACE, and 123 from Spain). Of these, 61 were identified as non-encapsulated pneumococci with a typical *lytA*-BsaAI-RFLP pattern (29 from Portugal and 32 from Spain) and 175 were identified as non *S. pneumoniae* with an atypical *lytA*-BsaAI-RFLP pattern (66 from Portugal, 22 from GRACE, and 87 from Spain). The 11 isolates reported here (four from Portugal, three from GRACE, and four from Spain) exhibited conflicting (or novel) results when presumptive identification based on optochin susceptibility and bile solubility was compared to the one suggested by the *lytA*-BsaAI-RFLP typing system.

Optochin susceptibility in CO_2 and in O_2 atmosphere. Optochin susceptibility was tested by disk diffusion using commercially available optochin discs (5 µg; 6 mm; Oxoid, Hampshire, England). Discs were applied to overnight cultures plated in blood

agar (trypticase soy agar supplemented with 5% defibrinated sheep blood). Plates were incubated overnight at 37°C in 5% CO_2 atmosphere or in ambient atmosphere as described by Arbique *et al.* (Arbique *et al.*, 2004). Isolates were considered resistant to optochin when they displayed inhibition zones smaller than 14 mm.

Bile solubility test. The bile solubility assay was performed according to standard procedures: colonies from an overnight culture were suspended in 1 mL of a 0.85% NaCl (w/v) solution to a turbidity equal to 0.5-1 McFarland standard (Rouff *et al*., 2003). This suspension was distributed into two tubes (500 µL each tube) and 200 µL of a 10% deoxycholate solution were added to one tube while the other received 200 µL of a 0.85% NaCl (w/v) solution (control). Both tubes were incubated at 37ºC for up to 2h. A sample was considered soluble in bile when clearing of the turbidity occurred in the tube with deoxycholate but not in the control.

Capsular typing. Capsular type assignment was performed by the Quellung reaction and by multiplex PCR as previously described (Brito *et al*., 2003; Pai *et al*., 2006; Simões *et al*., 2011).

***lytA*-BsaAI-RFLP signatures.** The entire *lytA* gene was amplified by PCR using primers previously described (LA5_Ext: 5'-AAGCTTTTTAGTCTGGGGTG-3' and LA3_Ext: 5'-AAGCTTTTTCAAGACCTAATAATATG-3'), yielding a PCR product of approximately 1200 bp (Obregon *et al.*, 2002). Typical (characteristic of pneumococcal *lytA*) or atypical (characteristic of non-pneumococcal *lytA*) RFLP signatures were determined as described before by digesting the PCR product with BsaAI and separating the fragments by agarose gel electrophoresis (Llull *et al.*, 2006).

Real-time PCR targeting *lytA and piaB*. The presence of the *lytA* gene was tested by real-time PCR using previously described primers (lytA_F: 5'-ACGCAATCTAG CAGATGAAGCA-3' and lytA_R: 5'-TCGTGCGTTTTAATTCCAGCT-3') and probe (5'-FAM-GCCGAAAACGCTTGATACAGGGAG-3'-BHQ1) (Carvalho Mda *et al.*, 2007). The presence of the *piaB* gene was tested by real-time PCR using previously described primers (piaF: 5'-CATTGGTGGCTTAGTAAGTGCAA-3' and piaR: 5'-TACTAAC ACAAGTTCCTGATAAGGCAAGT-3') and probe (5'-FAM-TGTAAGCGGAAAAGCAG GCCTTACCC-3'-BHQ1) (Trzcinski *et al.*, 2013). The assays were carried out in a final volume of 25 µL using the FastStart TaqMan Probe Master (Roche) containing 2.5 µL of 0.2 ng/µL DNA, 0.15 µM of each primer, and 0.075 µM of probe. The assay was performed three times on different days and DNA from *S. pneumoniae* TIGR4 (positive control) and *S. pseudopneumoniae* ATCC BAA-960 (negative control) was used in every run. DNA was amplified with CFX96 real-time system (Bio-Rad) with the cycling parameters previously described (Carvalho Mda *et al.*, 2007). Samples were considered positive when cycle threshold (Ct) values were equal to or below 35.

DNA sequencing. Sequencing reactions needed for the methods described below were conducted at Macrogen, Inc. (Amsterdam, The Netherlands). Sequencing analysis was done with DNAStar (Lasergene).

***lytA* sequencing analysis.** PCR products of approximately 1200 bp containing the entire *lytA* gene (957 bp) were obtained as described above (Obregon *et al.*, 2002). Sequencing was conducted at Macrogen and subsequent analysis of the sequences of the *lytA* gene was done with DNAstar. *lytA* sequences were also obtained for

strains TIGR4 (pneumococcus, NCBI accession number AE 005672.3) and ATCC BAA-960 (*S. pseudopneumoniae*, NCBI accession number AM113495.1), to be used for comparison. Nucleotide sequences of *lytA* gene described in this study were deposited at the GenBank database with the accession numbers KT253593-KT253603.

Multilocus sequence typing (MLST). Amplification of internal fragments of the seven housekeeping genes – *aroE*, *gdh*, *gki, recP*, *spi*, *xpt*, and *ddl* - was done according to the MLST scheme developed by Enright and Spratt for *S. pneumoniae* (Enright and Spratt, 1998). Sequencing analysis was done with DNAStar. Allele number assignment was done at the international MLST database for *S. pneumoniae* (www.mlst.net).

Multilocus sequence analysis (MLSA) for viridans group streptococci. Amplification of internal fragments of the seven housekeeping genes – *map*, *pfl*, *ppaC*, *pyk*, *rpoB*, *sodA*, and *tuf* – was done according to the scheme developed and validated for viridans group of streptococci by Bishop *et al*., except for the primer sodA-dn, that had an R to Y substitution (5'-A**Y**RTARTAM GCRTGYTCCCARACRTC-3') based on published sequences of strains TIGR4 (*S. pneumoniae*, NCBI accession number AE005672.3), B6 (*S. mitis*, NCBI accession number NC_013853.1), and IS7493 (*S. pseudopneumoniae*, NCBI accession number CP002925.1) (Bishop *et al.*, 2009). Phylogenetic analysis of the concatenated sequences of both strains analyzed in this study and the ones deposited at the eMLSA database (427 strains of the viridans group of *Streptococcus*; http://www.emlsa.net/) was performed using MEGA6.06 (http://www.mega software.net/): sequences were aligned by ClustalW using default

parameters (gap opening penalty of 15 and gap extension penalty of 6.66 for both pairwise and multiple alignment; IUB as DNA weight matrix, with a transition weight of 0.5). A minimum-evolution phylogenetic tree was constructed using default parameters (maximum composite likelihood was used as the substitution model of nucleotides, with transitions and transversions as the substitutions to include; uniform rates among sites; homogeneous pattern among lineages; complete deletion of gaps and missing data and close-neighbor-interchange was used as the ME heuristic method, based on an initial tree by neighbor-joining). Different concatenated sequences of the strains analyzed were arbitrarily named viridans MLSA profiles 1 to 10 and species assignment was inferred based on clustering analysis of the study isolates with the strains from the MLSA database.

Results

Phenotypic and genotypic characterization of the strains. Characteristics of the strains studied are described in Table 1, which also shows a summary of all results. All strains grew in gentamycin blood agar and displayed pneumococcus-like colony morphology albeit some were optochin resistant. The presence of *cpsA* or of other capsular genes (screened by PCR serotyping as described in the Material and Methods section) could not be detected in any of the strains, suggesting the absence of a pneumococcal capsule.

Atypical and novel *lytA*-BsaAI-RFLP patterns found in pneumococci. One pneumococcal strain (GRA218B) displayed an atypical *lytA*-BsaAI-RFLP pattern,

Table 1. Properties of the strains characterised in this study.

Strain	Clinical source	Country and year of isolation	Patient age (years) and gender	Optochin susceptibility		Bile solubility	*S. pneumoniae* MLST pattern (ST)[b]	*lytA*		*piaB* real-time PCR (Ct)	Viridans MLSA classification[d]	
				in 5% CO_2	in O_2			RFLP pattern[c]	Real-time PCR (Ct)		Profile	Species assignment
GRA036B	NP	SP, 2008	69, M	S	S	Pos	13, 8, 65, 1, 60, 16, 6 (508)	A	Pos (23)	Pos (24)	1	*S. pneumoniae*
Spain6220	Sputum	SP, 2002	71, M	S	S	Pos	13, 8, 65, 1, 60, 16, 6 (508)	A	Pos (24)	Pos (27)	2	*S. pneumoniae*
Spain7582	Sputum	SP, 2005	61, M	S	S	Pos	13, 8, 65, 1, 60, 16, 6 (508)	A	Pos (24)	Pos (28)	1	*S. pneumoniae*
GRA218B	NP	SP, 2010	29, F	S	S	Pos	8, 10, 84, 1, 2, 14, 59 (8073)	Atypical	Neg	Neg	3	*S. pneumoniae*
GRA254A	Sputum	BE, 2009	53, M	S	S	Pos	242(99%), 67(97%), 48(97%), 38(97%), 393(95%), 46(97%), nd	B	Neg	Neg	4	*S. mitis*
PT8543	NP	PT, 2011	4, M	R	R	Pos	106(98%), 414(98%), 318(96%), 195(97%), 193(96%), nd, 2(98%)	B	Neg	Neg	5	*S. mitis*
PT8638	NP	PT, 2011	5, M	R	R	Pos	59(97%), 94(98%), 3(96%), 195(97%), 40(95%), nd, 2(98%)	B	Neg	Neg	6	*S. mitis*
PT9018	NP	PT, 2012	4, M	R	R	Pos	242(98%), nd, 51(98%), 80(98%), 313(99%), 179(95%), 107(97%)	B	Neg	Neg	7	*S. mitis*
PT8238	NP	PT, 2011	3, M	R	R	Pos	88(95%), 94(98%), 318(97%), 67(96%), 182(96%), 153(94%), 545(96%)	C	Neg	Neg	8	*S. mitis*
Spain2270	CSF	SP, 2009	64, M	R	S	Pos	103(99%), 381(98%), 94(97%), 37(99%), 389(99%), nd, 447(97%)	Typical	Pos (24)	Neg	9	*S. pseudopneumoniae*
Spain9880	Sputum	SP, 2009	53, M	R	S	Pos	103(99%), 381(98%), 94(97%), 37(99%), 389(99%), nd, 447(97%)	Typical	Pos (24)	Neg	10	*S. pseudopneumoniae*

[a]NP, nasopharynx; CSF, cerebrospinal fluid; SP, Spain; BE, Belgium; PT, Portugal; S, susceptible; R, resistant; pos, positive; neg, negative.

[b]Allelic profile corresponds to the following genes: *aroE, gdh, gki, recP, spi, xpt,* and *ddl.* When a % is indicated it means that the allele of the strain is divergent from all the alleles described at the *S. pneumoniae* MLST database as of April 2015. The information indicates the similarity with the closest match.

[c]Typical, characteristic of pneumococcal *lytA;* atypical, characteristic of non-pneumococcal *lytA*; A, B and C, new patterns described here.

[d]Different concatenated sequences were arbitrarily named viridans MLSA profiles 1 to 10; and species assignment was inferred based on clustering analysis of the study isolates with the strains from the MLSA database (http://www.emlsa.net/).

usually associated with non-pneumococci (Fig. 1) (Llull *et al.*, 2006). This strain was both optochin susceptible and bile soluble, belonged to ST8073, and clustered with pneumococci by MLSA (Table 1, Fig. 2). Sequence analysis of the *lytA* gene confirmed a 98% similarity of the amino acid sequence with the LytA homologue of ATCC BAA-960, the *S. pseudopneumoniae* strain used as reference (Fig. S2).

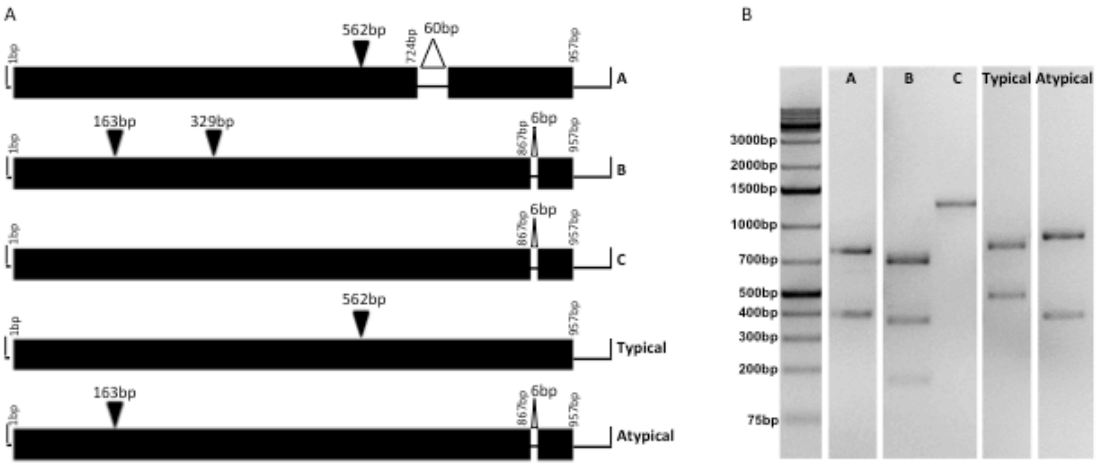


Figure 1. *lytA*-BsaAI-RFLP patterns.
(A) Schematic representation of the DNA fragments produced by digestion of *lytA* gene with BsaAI. Solid triangles show the restriction sites of BsaAI. Open triangles represent deletions. (B) Separation of BsaAI-digested fragments by agarose gel electrophoresis.

Three other pneumococcal strains (GRA036B, Spain6220, and Spain7582) displayed a *lytA*-BsaAI-RFLP pattern not previously described (pattern A, Fig. 1). These strains were both optochin susceptible and bile soluble. Also, these strains belonged to ST508 and clustered with pneumococci by MLSA (Table 1, Fig. 2). Such similarities between these strains suggest that they belong to a common lineage that may be disseminated. Sequence analysis of the *lytA* gene revealed a 60bp deletion at 724bp, resulting in the loss of 20 amino acids (KKIAEKWYYFDGEGAMKTGW). Apart from this deletion, the LytA amino acid sequence

shared a 99% similarity with that of TIGR4, the pneumococcal strain used as reference (Fig. S2).

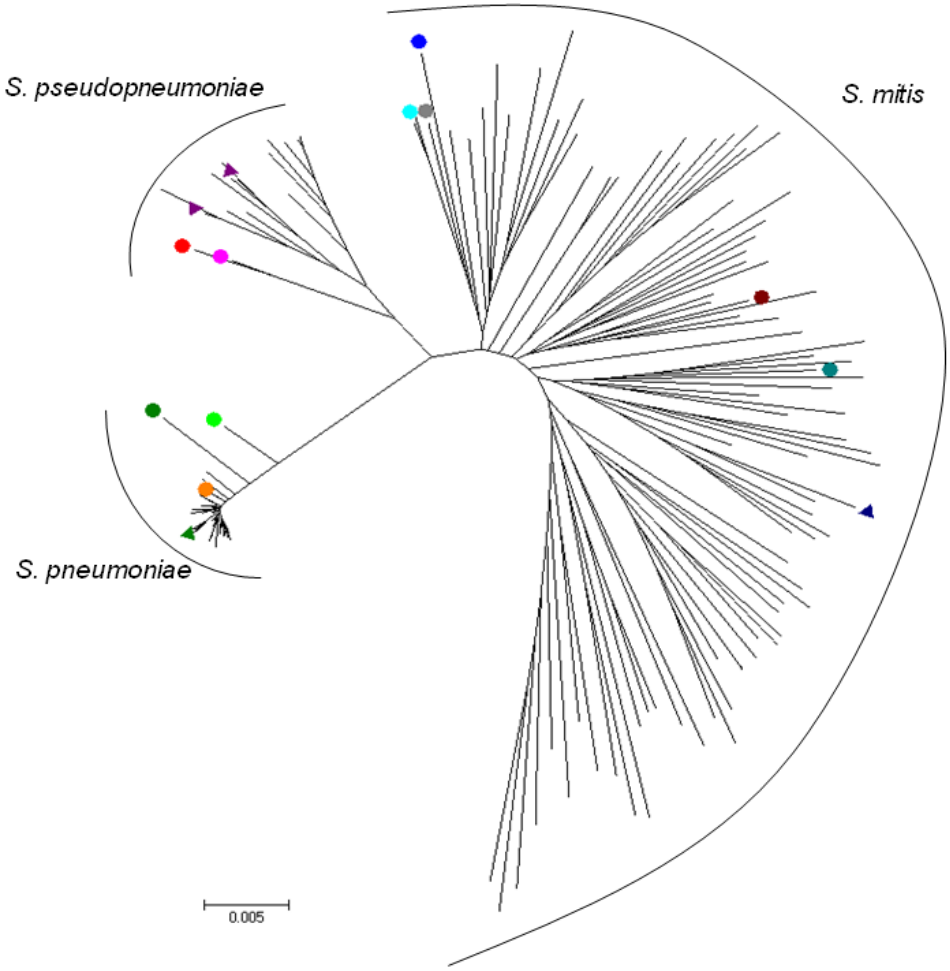


Figure 2. Genetic relationships of the strains analyzed in this study and *S. pneumoniae*, *S. pseudopneumoniae*, and *S. mitis* strains deposited at the eMLSA database.
Orange circle – GRA036B and Spain7582, light green circle – Spain6220, dark green circle – GRA218B, dark blue circle – GRA254A, blue-green circle – PT8543, maroon circle – PT8638, cyan circle – PT9018, gray circle – PT8238, rose circle – Spain 2270, red circle – Spain9880, green triangle – *S. pneumoniae* reference strain TIGR4, purple triangle – *S. pseudopneumoniae* reference strains ATCC-BAA 960 and IS7493, blue triangle – *S. mitis* reference strain DSM12643, no markers – *S. pneumoniae*, *S. pseudopneumoniae*, and *S. mitis* strains deposited at eMLSA database.

Typical and novel *lytA*-BsaAI-RFLP patterns found in non-pneumococci. Two non-pneumococcal strains (Spain2270 and Spain9880) displayed a typical *lytA*-BsaAI-RFLP pattern, usually associated with pneumococci (Fig. 1) (Llull *et al.*, 2006). These strains were optochin resistant in CO_2 atmosphere, optochin susceptible in aerobic conditions, and bile soluble. These strains could not be assigned to an ST according to the *S. pneumoniae* MLST database and clustered with *S. pseudopneumoniae* by MLSA (Table 1, Fig. 2). Such similarities between these strains suggest that they belong to a common lineage. Sequence analysis of the *lytA* gene confirmed a 99% similarity of the amino acid sequence with the LytA of TIGR4, the pneumococcal strain used as reference (Fig. S2).

Five other non-pneumococcal strains displayed two *lytA*-BsaAI-RFLP patterns not previously described (patterns B and C, Fig. 1). All these strains were bile soluble and while one was susceptible to optochin, the others were resistant. None of the strains could be assigned to an ST according to the *S. pneumoniae* MLST database and they all clustered with *S. mitis* by MLSA (Table 1, Fig. 2). Sequence analysis of the *lytA* gene of strains displaying pattern B (GRA254A, PT8543, PT8638, and PT9018) revealed a silent C329T mutation resulting in an extra BsaAI cutting site. On the other hand, sequence analysis of the *lytA* gene of the strain displaying pattern C (PT8238) revealed a silent C163T mutation resulting in the loss of the BsaAI cutting site. Apart from these mutations, the LytA amino acid sequences of these strains were 97-98% similar to that of the homologue of ATCC BAA-960, the *S. pseudopneumoniae* strain used as reference.

Misidentifications by real-time PCR targeting *lytA* and *piaB*. Pneumococcal identification by *lytA* real-time PCR failed for three of the 11 strains in this study: the pneumococcal strain

harboring the *S. pseudopneumoniae* homologue of the *lytA* gene (no amplification) and the two *S. pseudopneumoniae* strains encoding for (the pneumococcal) *lytA* (Ct 24 for both, Table 1). For all other cases described, the mutations found were located outside the annealing sites of the real-time PCR primers and probe and no misidentifications by real-time PCR were found (Fig. S2).

However, when pneumococcal identification by real-time PCR targeting *lytA* was complemented by real-time PCR targeting *piaB*, only the pneumococcal strain harboring the *S. pseudopneumoniae* homologue of the *lytA* gene remained misidentified, as this strain was also negative for *piaB* (Table 1).

Discussion

In this study we aimed to determine the genetic basis for the unexpected results given by *lytA*-BsaAI-RFLP and to investigate the accuracy of the *lytA* real-time PCR to classify 11 α-hemolytic streptococcal isolates displaying conflicting or novel results when characterized by the combination of optochin susceptibility, bile solubility, *lytA*-BsaAI-RFLP, and MLST. The main findings of our work were: (i) the identification of one pneumococcal carriage strain harboring a non-pneumococcal homologue of *lytA*; (ii) the identification of two invasive (cerebrospinal fluid and sputum) *S. pseudopneumoniae* strains harboring the characteristic pneumococcal *lytA*; and (iii) the misidentification of these three strains just referred to above by the commonly used *lytA* real-time PCR. In addition, novel *lytA*-BsaAI-RFLP patterns were identified and these were due to deletions or point mutations in *lytA*.

This is, to the best of our knowledge, the first report of a pneumococcal isolate harboring a non-pneumococcal homologue of *lytA.* Isolates with similar properties may have passed undetected in culture-independent studies relying solely on detection of *lytA* to identify pneumococci. On the other hand, non-pneumococcal isolates originating false positive results by the *lytA* real-time PCR strategy have been described (Carvalho Mda *et al.*, 2013). In particular, two copies of *lytA* (one copy of pneumococcal *lytA* and one copy of the non-pneumococcal homologue) were described in the genome of a clinical isolate (IS7493) of *S. pseudopneumoniae* (Shahinas *et al.*, 2011). However, *lytA* real-time PCR amplification of strain IS7493 did not retrieve a positive result (Tavares *et al*., unpublished).

The use of *piaB* real-time PCR to complement *lytA* real-time PCR has been proposed as *piaB* has been described as a pneumococcal specific non-ubiquitous gene that appears to be present in the majority of pneumococcal isolates (a notable exception are some non-capsulated pneumococci) (Tavares *et al.*, 2014; Trzcinski *et al.*, 2013; Whalan *et al.*, 2006). Whalan *et al*. found this gene in all encapsulated pneumococci tested (39 isolates covering 27 serotypes) and in six out of eight (75%) non-typeable pneumococci (Whalan *et al*., 2006). However, in a recent study of non-typeable pneumococci circulating in Portugal, we have only detected *piaB* in 12 out of 35 (34%) non-typeable pneumococci suggesting absence of this gene is common among these strains (Tavares *et al*., 2015). Nevertheless, the strategy of targeting *piaB* in addition to *lytA* clearly enhances the specificity of pneumococcal identification (Trzcinski *et al*., 2013). Even so, in the present study, one pneumococcal strain could not be identified based on these two real-time PCR assays.

The occurrence of genetic exchange between oral *Streptococcus* species has been well documented and horizontal gene transfer has been suggested as an important attenuator of putative species barriers (Chi *et al.*, 2007; Denapaite *et al.*, 2010; Donati *et al.*, 2010; Hanage *et al.*, 2005b; Johnston *et al.*, 2010). In fact, a smooth transition between pneumococcus species and its close relatives has been proposed (Hakenbeck *et al.*, 2001). On the other hand, Kilian *et al*. have proposed that both pneumococcus and *S. mitis* have evolved divergently from a pathogenic common ancestral: while pneumococcus maintained most of the ancestral virulence genes, *S. mitis* evolved to become a commensal (Kilian *et al.*, 2008; Kilian *et al.*, 2014).

It is important to highlight that, although exceptions have been reported here and elsewhere (Arbique *et al.*, 2004; Balsalobre *et al.*, 2006; Bosshard *et al.*, 2004; Nunes *et al.*, 2008; Sá-Leão *et al.*, 2006; Simões *et al.*, 2010; Whatmore *et al.*, 2000;), susceptibility to optochin, bile solubility, and assignment of a capsular type are, each one of them, excellent presumptive methods to identify the majority of pneumococcal isolates. In this report, all but one *S. mitis* strain were correctly identified based on optochin susceptibility. For dubious cases, the assignment of specific *lytA*-BsaAI-RFLP signatures and a multiplex PCR strategy have been proposed (Llull *et al.*, 2006; Simões *et al.*, 2011). However, this study suggests that with these methods, although rarely, misidentification can still occur. The assignment of a sequence type by the *S. pneumoniae* MLST scheme or the viridans MLSA scheme, although very useful as tools for species identification, are time-consuming and expensive for routine laboratories, thus often being used only in selected cases (Bishop *et al.*, 2009; Hanage *et al.*, 2005a). Alternatively, the determination of a pneumococcal-specific

sequence signature of 16S rRNA has also been proposed as an inexpensive identification tool (Scholz *et al.*, 2012).

One possible limitation of our study is that we did not systematically study the frequency at which isolates with the new or conflicting characteristics described in this study occur in collections of pneumococcal clinical and carriage isolates. This aim was beyond of scope of this study. Although all isolates were characterized by optochin susceptibility and bile solubility, *lytA*-BsaAI-RFLP was applied only with isolates presumptively identified as pneumococci and for which a serotype could not be assigned. Still, based on our experience, the frequency of such isolates appears to be low. In particular, among the colonization isolates from Portugal (n=1226) we have systematically performed the *lytA*-BsaAI-RFLP assay for presumptive pneumococcal isolates for which a capsular type could not be assigned (n=99). Of these, four *S. mitis* isolates (described in this study), corresponding to c.a. 4.0% of the tested samples, had unusual *lytA* patterns. In addition, in our collections the unusual pattern A was found in three of 61 (4.9%) *S. pneumoniae* isolates tested by the *lytA*-BsaAI-RFLP.

The future is heading towards automated screenings of unprocessed samples. Currently, *lytA* real-time PCR is the culture-independent methodology of choice and the association with a second gene such as *piaB* as proposed by Trzcinski *et al*. is an interesting strategy (Carvalho Mda *et al.*, 2007; Satzke *et al.*, 2013; Trzcinski *et al.*, 2013). Whole genome sequencing (WGS) is being increasingly used and an approach combining automated extraction of WGS information with MLST-extended schemes will undoubtedly reveal itself

very useful for the unambiguous classification of strains as the ones described in this study (Sabat *et al.*, 2013).

In conclusion, identification of pneumococci based on *lytA* detection, including real-time PCR, may lead to false results. This is of particular relevance in the increasingly frequent colonization studies relying solely on culture-independent methods targeting *lytA*.

Acknowledgments

This work was funded by Fundação para a Ciência e a Tecnologia, Portugal, through grants PTDC/BIA-MIC/64010/2006 and PTDC/BIA-BEC/098289/2008 to RSL, SFRH/BD/70147/2010 to DAT, SFRH/BD/27325/2006 to ASS, and Pest-OE/EQB/LAO004/2011 to Laboratório Associado de Oeiras. The funding agency had no involvement in the study design, collection, analysis, and interpretation of data, writing of the article, nor in the decision to submit the study for publication. We thank Sónia T. Almeida for assistance with MLST.

References

Arbique, J.C., C. Poyart, P. Trieu-Cuot, G. Quesne, G. Carvalho Mda, A.G. Steigerwalt, R.E. Morey, D. Jackson, R.J. Davidson and R.R. Facklam. (2004). Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. *J Clin Microbiol* **42**, 4686-96.

Balsalobre, L., A. Hernandez-Madrid, D. Llull, A.J. Martin-Galiano, E. Garcia, A. Fenoll and A.G. de la Campa. (2006). Molecular characterization of disease-associated streptococci of the mitis group that are optochin susceptible. *J Clin Microbiol* **44**, 4163-71.

Bishop, C.J., D.M. Aanensen, G.E. Jordan, M. Kilian, W.P. Hanage and B.G. Spratt. (2009). Assigning strains to bacterial species via the internet. *BMC Biol* **7**, 3.

Bochud, P.Y., T. Calandra and P. Francioli. (1994). Bacteremia due to viridans streptococci in neutropenic patients: a review. *Am J Med* **97**, 256-64.

Bosshard, P.P., S. Abels, M. Altwegg, E.C. Bottger and R. Zbinden. (2004). Comparison of conventional and molecular methods for identification of aerobic catalase-negative gram-positive cocci in the clinical laboratory. *J Clin Microbiol* **42**, 2065-73.

Brito, D.A., M. Ramirez and H. de Lencastre. (2003). Serotyping *Streptococcus pneumoniae* by multiplex PCR. *J Clin Microbiol* **41**, 2378-84.

Carvalho Mda, G., M.L. Tondella, K. McCaustland, L. Weidlich, L. McGee, L.W. Mayer, A. Steigerwalt, M. Whaley, R.R. Facklam, B. Fields, G. Carlone, E.W. Ades, R. Dagan and J.S. Sampson. (2007). Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *J Clin Microbiol* **45**, 2460-6.

Carvalho Mda, G., F.C. Pimenta, I. Moura, A. Roundtree, R.E. Gertz, Jr., Z. Li, G. Jagero, G. Bigogo, M. Junghae, L. Conklin, D.R. Feikin, R.F. Breiman, C.G. Whitney and B.W. Beall. (2013). Non-pneumococcal mitis-group streptococci confound detection of pneumococcal capsular serotype-specific loci in upper respiratory tract. *PeerJ* **1**, e97.

Chi, F., O. Nolte, C. Bergmann, M. Ip and R. Hakenbeck. (2007). Crossing the barrier: evolution and spread of a major class of mosaic pbp2x in *Streptococcus pneumoniae*, *S. mitis* and *S. oralis*. *Int J Med Microbiol* **297**, 503-12.

Denapaite, D., R. Bruckner, M. Nuhn, P. Reichmann, B. Henrich, P. Maurer, Y. Schahle, P. Selbmann, W. Zimmermann, R. Wambutt and R. Hakenbeck. (2010). The genome of *Streptococcus mitis* B6-what is a commensal? *PLoS One* **5**, e9426.

Donati, C., N.L. Hiller, H. Tettelin, A. Muzzi, N.J. Croucher, S.V. Angiuoli, M. Oggioni, J.C. Dunning Hotopp, F.Z. Hu, D.R. Riley, A. Covacci, T.J. Mitchell, S.D. Bentley, M. Kilian, G.D. Ehrlich, R. Rappuoli, E.R. Moxon and V. Masignani. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**, R107.

Douglas, C.W., J. Heath, K.K. Hampton and F.E. Preston. (1993). Identity of viridans streptococci isolated from cases of infective endocarditis. *J Med Microbiol* **39**, 179-82.

Enright, M.C. and B.G. Spratt. (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* **144 (Pt 11)**, 3049-60.

Hakenbeck, R., N. Balmelle, B. Weber, C. Gardes, W. Keck and A. de Saizieu. (2001). Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. *Infect Immun* **69**, 2477-86.

Hanage, W.P., T. Kaijalainen, E. Herva, A. Saukkoriipi, R. Syrjanen and B.G. Spratt. (2005a). Using multilocus sequence data to define the pneumococcus. *J Bacteriol* **187**, 6223-30.

Hanage, W.P., C. Fraser and B.G. Spratt. (2005b). Fuzzy species among recombinogenic bacteria. *BMC Biol* **3**, 6.

Ioannidou, S., P.T. Tassios, A. Kotsovili-Tseleni, M. Foustoukou, N.J. Legakis and A. Vatopoulos. (2001). Antibiotic resistance rates and macrolide resistance phenotypes of viridans group streptococci from the oropharynx of healthy Greek children. *Int J Antimicrob Agents* **17**, 195-201.

Johnston, C., J. Hinds, A. Smith, M. van der Linden, J. Van Eldere and T.J. Mitchell. (2010). Detection of large numbers of pneumococcal virulence genes in streptococci of the mitis group. *J Clin Microbiol* **48**, 2762-9.

Keith, E.R., R.G. Podmore, T.P. Anderson and D.R. Murdoch. (2006). Characteristics of *Streptococcus pseudopneumoniae* isolated from purulent sputum samples. *J Clin Microbiol* **44**, 923-7.

Kilian, M., K. Poulsen, T. Blomqvist, L.S. Havarstein, M. Bek-Thomsen, H. Tettelin and U.B. Sorensen. (2008). Evolution of *Streptococcus pneumoniae* and its close commensal relatives. *PLoS One* **3**, e2683.

Kilian, M., D.R. Riley, A. Jensen, H. Bruggemann and H. Tettelin. (2014). Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. *MBio* **5**, e01490-14.

Llull, D., R. Lopez and E. Garcia. (2006). Characteristic signatures of the *lytA* gene provide a basis for rapid and reliable diagnosis of *Streptococcus pneumoniae* infections. *J Clin Microbiol* **44**, 1250-6.

Messmer, T.O., C.G. Whitney and B.S. Fields. (1997). Use of polymerase chain reaction to identify pneumococcal infection associated with hemorrhage and shock in two previously healthy young children. *Clin Chem* **43**, 930-5.

Nunes, S., R. Sá-Leão and H. de Lencastre. (2008). Optochin resistance among *Streptococcus pneumoniae* strains colonizing healthy children in Portugal. *J Clin Microbiol* **46**, 321-4.

Obregon, V., P. Garcia, E. Garcia, A. Fenoll, R. Lopez and J.L. Garcia. (2002). Molecular peculiarities of the *lytA* gene isolated from clinical pneumococcal strains that are bile insoluble. *J Clin Microbiol* **40**, 2545-54.

Pai, R., R.E. Gertz and B. Beall. (2006). Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates. *J Clin Microbiol* **44**, 124-31.

Rolo, D., S.S. A, A. Domenech, A. Fenoll, J. Linares, H. de Lencastre, C. Ardanuy and R. Sá-Leão. (2013). Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. *PLoS One* **8**, e57047.

Romero, P., R. Lopez and E. Garcia. (2004). Characterization of LytA-like N-acetylmuramoyl-L-alanine amidases from two new *Streptococcus mitis* bacteriophages provides insights into the properties of the major pneumococcal autolysin. *J Bacteriol* **186**, 8229-39.

Rouff, K., R.A. Whiley, D. Beighton. (2003). *Streptococcus.* In Murray, P.R., E.J. Barron. J.H. Jorgensen, M.A. Pfaller and R.H. Yolken (Eds, *Manual of Clinical Microbiology* (8th ed., pp. 405-421). Washington, D.C.: American Society for Microbiology).

Sabat, A.J., A. Budimir, D. Nashev, R. Sá-Leão, J. van Dijl, F. Laurent, H. Grundmann, A.W. Friedrich and E.S.G.o.E. Markers. (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill* **18**, 20380.

Sá-Leão, R., A.S. Simões, S. Nunes, N.G. Sousa, N. Frazão and H. de Lencastre. (2006). Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. *Microbiology* **152**, 367-76.

Satzke, C., P. Turner, A. Virolainen-Julkunen, P.V. Adrian, M. Antonio, K.M. Hare, A.M. Henao-Restrepo, A.J. Leach, K.P. Klugman, B.D. Porter, R. Sa-Leao, J.A. Scott, H. Nohynek, K.L. O'Brien and W.H.O.P.C.W. Group. (2013). Standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*: updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. *Vaccine* **32**, 165-79.

Scholz, C.F., K. Poulsen and M. Kilian. (2012). Novel molecular method for identification of *Streptococcus pneumoniae* applicable to clinical microbiology and 16S rRNA sequence-based microbiome studies. *J Clin Microbiol* **50**, 1968-73.

Shahinas, D., G.S. Tamber, G. Arya, A. Wong, R. Lau, F. Jamieson, J.H. Ma, D.C. Alexander, D.E. Low and D.R. Pillai. (2011). Whole-genome sequence of *Streptococcus pseudopneumoniae* isolate IS7493. *J Bacteriol* **193**, 6102-3.

Simões, A.S., R. Sá-Leão, M.J. Eleveld, D.A. Tavares, J.A. Carriço, H.J. Bootsma and P.W. Hermans. (2010). Highly penicillin-resistant multidrug-resistant pneumococcus-like strains colonizing children in Oeiras, Portugal: genomic characteristics and implications for surveillance. *J Clin Microbiol* **48**, 238-46.

Simões, A.S., C. Valente, H. de Lencastre and R. Sá-Leão. (2011). Rapid identification of noncapsulated *Streptococcus pneumoniae* in nasopharyngeal samples allowing detection of co-colonization and reevaluation of prevalence. *Diagn Microbiol Infect Dis* **71**, 208-16.

Tavares, D.A., A.S. Simões, H.J. Bootsma, P.W. Hermans, H. de Lencastre and R. Sá-Leão. (2014). Non-typeable pneumococci circulating in Portugal are of *cps* type NCC2 and have genomic features typical of encapsulated isolates. *BMC Genomics* **15**, 863.

Trzcinski, K., D. Bogaert, A. Wyllie, M.L. Chu, A. van der Ende, J.P. Bruin, G. van den Dobbelsteen, R.H. Veenhoven and E.A. Sanders. (2013). Superiority of trans-oral over trans-nasal sampling in detecting *Streptococcus pneumoniae* colonization in adults. *PLoS One* **8**, e60520.

Wester, C.W., D. Ariga, C. Nathan, T.W. Rice, J. Pulvirenti, R. Patel, F. Kocka, J. Ortiz and R.A. Weinstein. (2002). Possible overestimation of penicillin resistant *Streptococcus pneumoniae* colonization rates due to misidentification of oropharyngeal streptococci. *Diagn Microbiol Infect Dis* **42**, 263-8.

Whalan, R.H., S.G. Funnell, L.D. Bowler, M.J. Hudson, A. Robinson and C.G. Dowson. (2006). Distribution and genetic diversity of the ABC transporter lipoproteins PiuA and PiaA within *Streptococcus pneumoniae* and related streptococci. *J Bacteriol* **188**, 1031-8.

Whatmore, A.M., A. Efstratiou, A.P. Pickerill, K. Broughton, G. Woodard, D. Sturgeon, R. George and C.G. Dowson. (2000). Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "Atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. *Infect Immun* **68**, 1374-82.

Supporting information

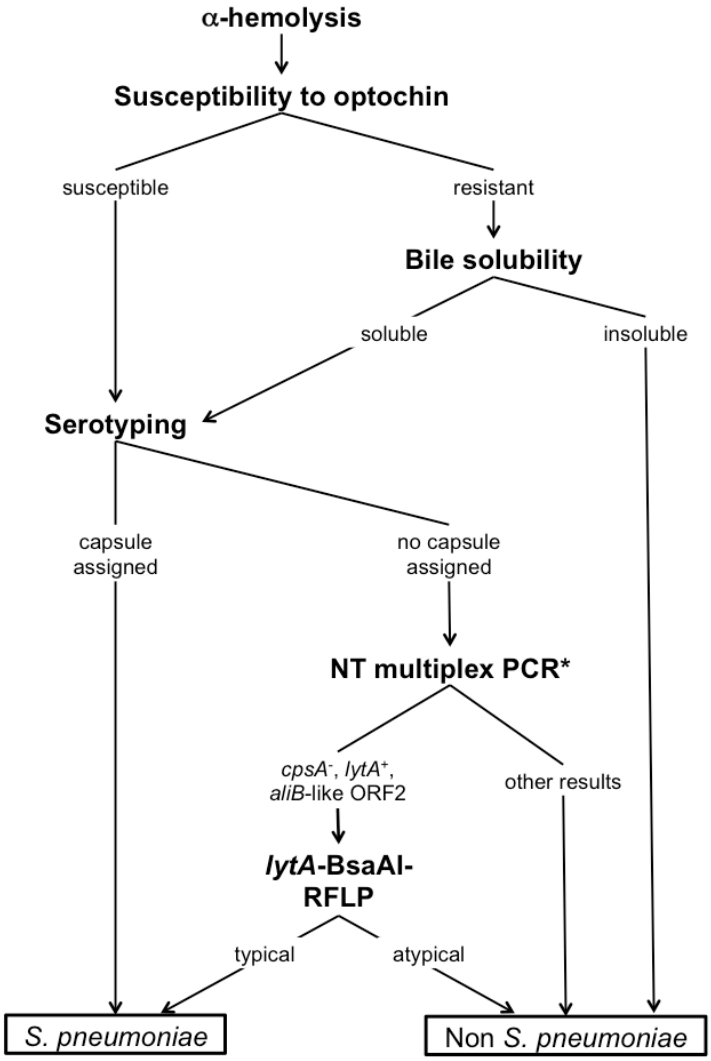
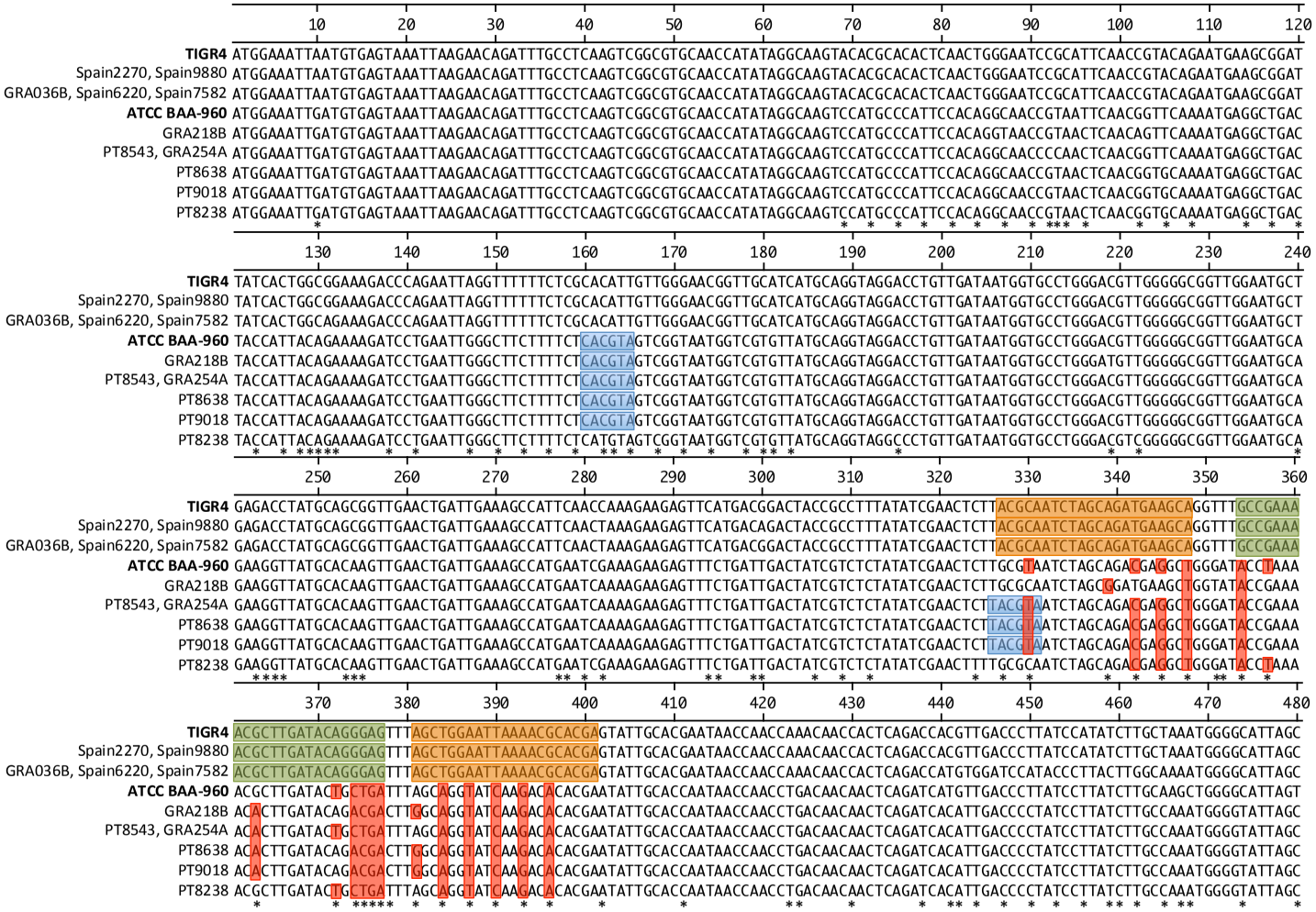


Figure S1. Routine laboratory workflow for the identification of *S. pneumoniae*.
*As described in (Simões *et al.*, 2010).


TIGR4
Spain2270, Spain9880
GRA036B, Spain6220, Spain7582
ATCC BAA-960
GRA218B
PT8543, GRA254A
PT8638
PT9018
PT8238

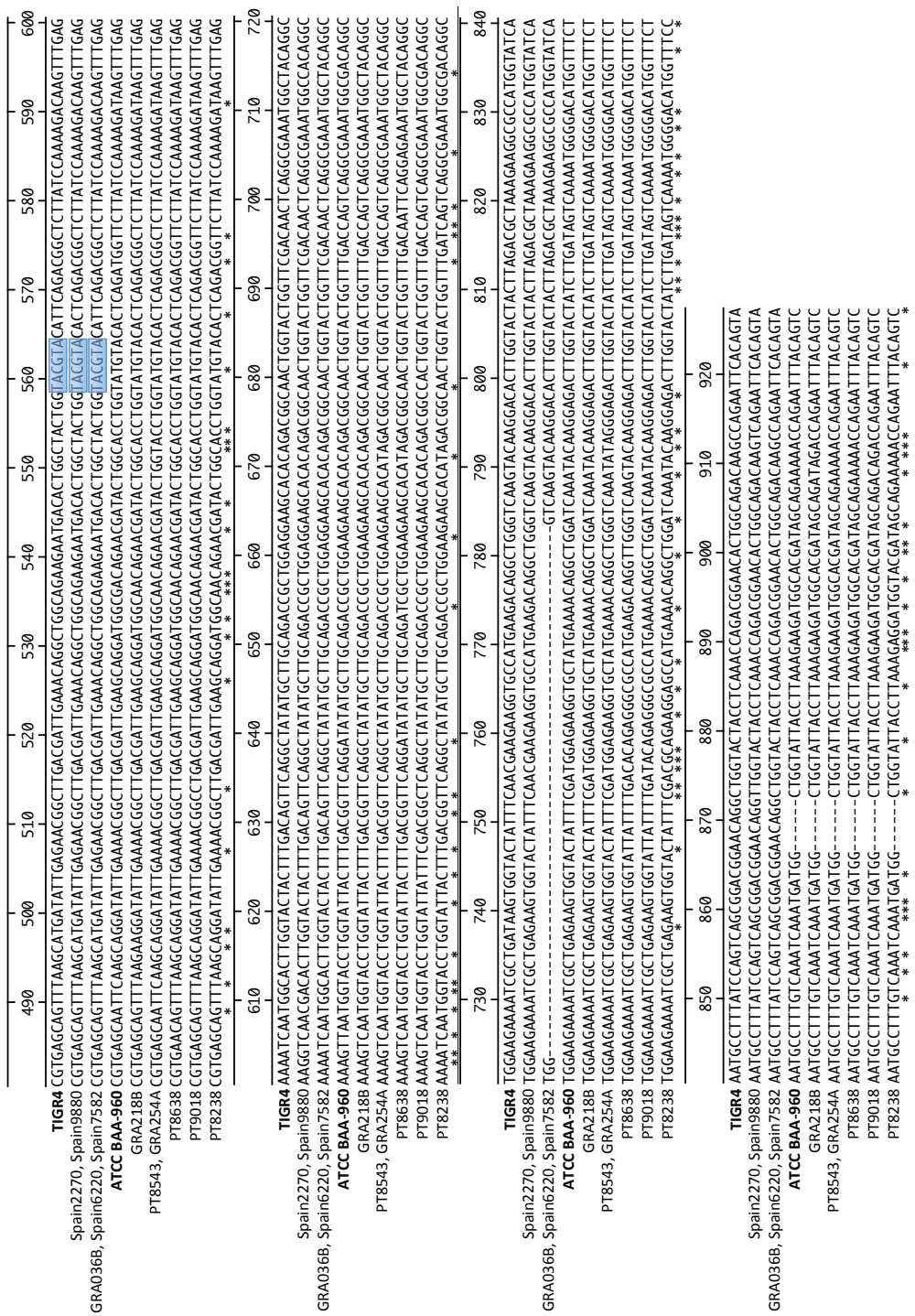


Figure S2. *lytA* sequences of the strains analyzed in this study.

In bold are the *lytA* sequences used as control (*S. pseudopneumoniae* ATCC BAA-960 and *S. pneumoniae* TIGR4); blue shadow indicates the restriction sites for BsaAI; green and orange shadows indicate probe and primers annealing sites (respectively) of the *lytA* real-time PCR assay; * indicates base substitutions; red shadow indicates substitutions in the annealing regions (primers and probe) of the *lytA* real-time PCR assay.

Chapter 4

Improved identification of *Streptococcus pneumoniae* by a real-time PCR assay targeting SP_2020

In preparation: D. A. Tavares, R. J. Carvalho, S. Handem, H. de Lencastre, J. Hinds, and R. Sá-Leão. **Improved identification of *Streptococcus pneumoniae* by a real-time PCR assay targeting SP_2020.**

Contributions:

D. A. Tavares was responsible for all experimental work, with the exception of design of *hylA* primers and probes and DNA isolation of most strains, which was performed by R. J. Carvalho, and sequencing of multilocus sequence analysis alleles, which was performed by S. Handem. Isolates were previously characterised by optochin susceptibility, bile solubility, serotyping, and *lytA*-BsaAI-RFLP under the scope of other studies.

Summary

In 2013, the *lytA*-CDC real-time PCR assay was recommended by the WHO as the gold-standard culture-independent assay for the identification of pneumococcus. In addition, a *piaB* assay has been proposed by others to be used in parallel to increase the specificity of the identification. In this study we aimed to evaluate the performance of *lytA*-CDC and *piaB* assays for the identification of pneumococcus and to design and evaluate the performance of two new assays targeting *hylA* and SP_2020. For that, a collection of 150 pneumococci (of 50 capsular types plus non-encapsulates pneumococcus) and 433 non-pneumococci (including type strains of 23 *Streptococcus* species) was tested. SP_2020 and *lytA*-CDC were the assays with the best performance (100% sensitivity and 99.3% and 98.7% positive predictive value (PPV), respectively, p=0.564), followed by *hylA* (100% sensitivity and 83.8% PPV), and *piaB* (93.3% sensitivity and 98.6% PPV). All assays misidentified strains of other *Streptococcus* species as pneumococcus. In particular, *lytA*-CDC misidentified two *S. pseudopneumoniae* strains as pneumococcus. Moreover, *piaB* misidentified pneumococcal strains of serotypes 6B, 11A, 23F, and non-encapsulated as non-pneumococcus. According to our results, SP_2020 in combination with *lytA*-CDC should be considered as an alternative to the *lytA*-CDC or the *lytA*-CDC and *piaB* combination, especially for culture-independent identification of pneumococcus in polymicrobial samples.

Introduction

The identification of the human pathogen *Streptococcus pneumoniae* (pneumococcus) may be a demanding task. In fact, exceptions to the traditional phenotypic assays (susceptibility to optochin, cell wall lysis by sodium deoxycholate (bile solubility), and assignment of a capsular type by serotyping) have all been described (Mundy *et al.*, 1998; Whatmore *et al.*, 2000; Arbique *et al.*, 2004). Also, molecular assays are often hampered by the frequent genetic exchange between pneumococcus and other species of the viridans group streptococci, mainly *S. pseudopneumoniae* and *S. mitis* (Whatmore *et al.*, 2000). In carriage studies investigating vaccine efficacy and resistance to antibiotics, it has been previously shown that misidentification of closely related species as pneumococci may lead to falsely increased rates of resistance (Wester *et al.*, 2002; Richter *et al.*, 2008; Simões *et al.*, 2010).

Although the WHO recommended algorithm for the identification of pneumococcus still relies on optochin susceptibility, bile solubility, and serotyping of cultured α-haemolytic colonies, the method of choice for culture-independent assays is a real-time PCR assay targeting the gene *lytA* developed by Carvalho *et al.* (*lytA*-CDC) (Carvalho Mda *et al.*, 2007; Satzke *et al.*, 2013). LytA is the major autolysin of pneumococcus and has been described as ubiquitous and specific of this species (Pozzi *et al.*, 1989). The performance of this real-time PCR assay was initially tested with a collection of 67 *S. pneumoniae* and 104 non-pneumococcal isolates. The latter group included 13 viridans group streptococci not identified to the species level. This

method has been extensively used by different laboratories in both disease and carriage studies (Trzcinski *et al.*, 2013; del Amo *et al.*, 2015).

In addition, a second real-time PCR assay, targeting *piaB*, a permease of an ABC transporter involved in iron uptake and virulence, has been used by others to increase the specificity of pneumococcal identification by this method (Trzcinski *et al.*, 2013; Wyllie *et al.*, 2014; Brown *et al.*, 2001). Although this system has been described as pneumococcus-specific, it is not ubiquitous, being absent from some non-encapsulated pneumococci (non-typeable) (Whalan *et al.*, 2006; Tavares *et al.*, 2014).

Albeit there is some evidence that homologues of *lytA* (and of other pneumococcal genes) can be present in closely related species of the viridans group of *Streptococcus*, to the best of our knowledge, this has not been sufficiently tested. Therefore, it is not clear how the potential presence of these genetic determinants in non-pneumococcal isolates could affect the performance of the currently used *lytA*-CDC and *piaB* real-time PCR assays in the identification of pneumococcus.

In this study we aimed to evaluate the performance of four real-time PCR assays for the identification of pneumococcus using a large collection of α-haemolytic non-pneumococcal isolates (n=402). Two control collections of 150 pneumococcal strains (of 50 serotypes plus non-typeables) and 31 type strains from 23 other *Streptococcus* species were also tested. The four real-time PCR assays evaluated were the *lytA*-CDC and *piaB* assays previously described and two novel assays targeting *hylA* (hyaluronidase) and SP_2020 (putative transcriptional regulator).

Materials and methods

Study collections. Three collections were tested: a *S. pneumoniae* control collection (n=150), a non-*S. pneumoniae spp.* control collection (n=31), and a test collection (n=402).

The *S. pneumoniae* control collection included 150 pneumococcal strains belonging to 50 serotypes plus non-typeables (NT). The represented serotypes were 1, 3, 4, 5, 6A, 6B, 7A, 7F, 8, 9A, 9L, 9N, 9V, 10A, 11A, 12A, 12B, 12F, 14, 15A, 15B, 15C, 15F, 16F, 17, 18A, 18B, 18C, 18F, 19A, 19F, 20, 21, 22F, 23A, 23B, 23F, 24B, 24F, 29, 31, 33B, 33F, 34, 35B, 35F, 37, 38, 39, and 42. This collection included the prototype strains of 27 Pneumococcal Molecular Epidemiology Network (PMEN) clones (http://www.sph.emory.edu/PMEN/index.htm): Spain^{23F}-1, Spain^{6B}-2, Spain^{9V}-3, Tennessee^{23F}-4, Spain^{14}-5, Hungary^{19A}-6, S.Africa^{19A}-7, S.Africa^{6B}-8, England^{14}-9, CSR^{14}-10, CSR^{19A}-11, Finland^{6B}-12, S.Africa^{19A}, Taiwan^{19F}, Taiwan^{23F}-15, Poland^{23F}-16, Maryland^{6B}-17, Tennessee^{14}-18, Colombia^{5}-19, Poland^{6B}-20, Portugal^{19F}-21, Greece^{6B}-22, N. Carolina^{6A}-23, Utah^{35B}-24, Sweden^{15A}-25, Colombia^{23F}-26, and Portugal^{6A}-41.

The non-*S. pneumoniae spp.* control collection included 31 type strains of the following 23 *Streptococcus* species: *S. pseudopneumoniae* (PT5479, IS7943), *S. mitis* (DSM12643), *S. oralis* (DSM20066, DSM20379, DSM20395, DSM20627), *S. cristatus* (DSM8249), *S. gordonii* (DSM6777, DSM20568), *S. infantis* (DSM12492), *S. parasanguinis* (DSM6778), *S. peroris* (DSM12493), *S. sanguinis* (DSM20567), *S. sinensis* (DSM14990), *S. anginosus* (DSMZ20563), *S. constellatus* (NCTC11325), *S. intermedius* (NCTC11324), *S. salivarius* (DSM20560), *S. vestibularis* (DSM5636), viridans streptococci (DSM20377, DSM20392), *S. agalactiae* (DSMZ6784), *S. bovis* (DSM20480), *S. canis* (DSM20715), *S. dysgalactiae sub. dysgalactiae*

(DSMZ20662), *S. dysgalactiae sub. equisimilis* (DSMZ6176), *S. equi sub. zooepidemicus* (DSMZ20727), *S. equinus* (NCTC10389), *S. mutans* (DSM20523), *S. pyogenes* (DSM20565).

The test collection included 402 α-haemolytic non-pneumococcal isolates. The isolates were recovered between 1991 and 2012 from different carriage and disease studies and all belong to our in-house collection. Isolates were initially isolated based on the observation of α-haemolysis and colony morphology suggestive of pneumococcus, but were found to be of other streptococcal species when a combination of methods was applied (optochin susceptibility, bile solubility, serotyping, and *lytA*-BsaI-RFLP) (Sá-Leão *et al.*, 2009; Simões *et al.*, 2011; Llull *et al.*, 2006). Of the 402 α-haemolytic non-pneumococcal isolates tested, 346 were resistant to optochin, 25 were susceptible to optochin but bile insoluble, and 31 were susceptible to optochin and bile soluble but could not be assigned to a serotype. These latter 31 isolates were confirmed not to be pneumococcus by a multiplex PCR scheme previously described (Simões *et al.*, 2011) and the identification of characteristic non-pneumococcal *lytA*-BsaAI-RFLP signatures (Llull *et al.*, 2006).

DNA isolation. Genomic DNA was obtained by using the DNeasy Blood & Tissue kit (Qiagen) or MagNa Pure Compact Nucleic Acid Isolation kit (Roche Diagnostics GmbH) as recommended by the manufacturers. All DNA samples were diluted to 0.2ng/μL with sterile water.

Design of real-time PCR assay targeting *hylA* and SP_2020. Two novel targets, *hylA* and SP_2020, were tested. The *hylA* gene encodes for hyaluronidase and is part of

the core genome of pneumococcus proposed by Obert *et al.* (Obert *et al.*, 2006). Hyaluronidase seems to favour colonisation and dissemination of pneumococcus by contributing to pneumolysin-mediated damage of the epithelium (Feldman *et al.*, 2007). SP_2020 is a putative transcriptional regulator and, like *hylA*, belongs to the core genome of pneumococcus proposed by Obert *et al.* (Obert *et al.*, 2006).

To design the real-time PCR assay targeting *hylA*, the nucleotide sequence of the TIGR4 *hylA* gene (SP_0314) was blasted against the NCBI database (as of May 2013). Of 61 hints, 23 were identical pneumococcal nucleotide sequences and thus only one was considered for comparison. The nucleotide sequence of SP_0314 was then blasted against the remaining 39 hints (33 pneumococcus, three *S. constellatus*, two *S. intermedius*, and one *S. anginosus*) to identify a region highly conserved between pneumococci and highly variable amongst the other *Streptococcus* species identified. A 146bp region closer to the 3' region of SP_0314 was identified and one set of primers and one Cy5-labeled probe were designed: hylA_F (5'-CCAAATTAAACCAGGAATTGGA-3'), hylA_R (5'-CTCCAATTTCCGATAAGTGGCA-3'), and hylA_P (5'-Cy5-TCAAGTCAGGCGGACCGCA-BBQ-3').

To design the real-time PCR assay targeting SP_2020, the nucleotide sequence of the TIGR4 SP_2020 gene was blasted against the NCBI database (as of November 2015). Homology was found only to pneumococcal nucleotide sequences (29 sequences, 99-100% nucleotide similarity) and not to any other *Streptococcus* species. One set of primers and a FAM-labelled probe were designed: SP_2020_F (5'-TAAACAGTTTGCCTGTAGTCG-3'), SP_2020_R (5'-CCCGGATATCTCTTTCTGGA-3'), and SP_2020_P (5'-Fam-AACCTTTGTTCTCTCTCGTGGCAGCTCAA-BBQ-3').

The real-time PCR assays were tested and optimised for *S. pneumoniae* TIGR4 (NCBI accession number AE 005672.3) and *S. pseudopneumoniae* ATCC BAA-960 (NCBI accession number AM113495.1).

Real-time PCR targeting *lytA*, *piaB*, *hylA*, and SP_2020. The presence of the genes *lytA* and *piaB* was tested by using primers and probes previously described (Carvalho Mda *et al.*, 2007; Trzcinski *et al.*, 2013). The presence of the genes *hylA* and SP_2020 was tested by using primers and probes described above. For all reactions, 0.5ng of DNA were tested in a final volume of 25µL containing 1x FastStart TaqMan Probe Master (Roche), 0.15µM each primer, 0.075µM probe. DNA was amplified with the CFX96 Real-Time System Amplification (Bio-Rad) by using the following cycling conditions: 95°C for 10min followed by 45 cycles of 95°C for 15sec, 60°C (for *lytA*-CDC and *piaB*) or 55°C (for *hylA* and SP_2020) for 1min. Fluorescence was read after each of the 45 cycles.

All strains from *S. pneumoniae* control collection and non-*S. pneumoniae spp.* control collection were tested twice in different days. All strains from test collection were tested once, except when amplification occurred. In such cases, isolates were re-tested for confirmation. In addition, for each assay, a random selection of 10% of the strains from test collection was also independently selected for re-testing. DNA from *S. pneumoniae* TIGR4 (positive control) and *S. pseudopneumoniae* ATCC BAA-960 (negative control) was used in every run. Samples were considered positive when the cycle threshold (C_T) value was equal or below 35.

Performance of the real-time PCR assays for the identification of pneumococcus. To evaluate the performance of the real-time PCR assays for the identification of

pneumococcus, two parameters were estimated: the sensitivity, to estimate the percentage of pneumococci correctly identified among all the pneumococci, and the positive predictive value (PPV), to estimate the percentage of pneumococci among the isolates giving a positive result for a given assay. To compare PPVs of the different real-time PCR assays, the generalised score statistic test proposed by Leisenring *et al*. was used (Leisenring *et al.*, 2000).

Multilocus sequence analysis (MLSA) for viridans group streptococci (viridans MLSA). Amplification of internal fragments of the seven housekeeping genes *map*, *pfl*, *ppaC*, *pyk*, *rpoB*, *sodA*, and *tuf* was done as described (Bishop *et al.*, 2009; Simões *et al.*, 2016). Sequencing reactions were conducted at Macrogen, Inc and sequence analysis was performed using DNAstar (Lasergene). Phylogenetic analysis of the concatenated sequences in comparison with the eMLSA database (http://www.emlsa.net/) was performed using MEGA6.06 (http://www.mega software.net) as described: alignment of sequences by ClustalW and construction of a minimum-evolution phylogenetic tree using default parameters (Simões *et al.*, 2016).

Results and discussion

To evaluate the sensitivity of the real-time PCR assays, 150 pneumococcal strains were used. For all pneumococcal strains, a positive real-time PCR result was obtained for *lytA*, *hylA*, and SP_2020 (Figure 1); in addition, 140 (93.3%) strains were positive for *piaB*. The 10 strains that had a negative result for *piaB* were of capsular

types NT (n=6), 23F (n=2), 6B (n=1), and 11A (n=1). Hence, by using this collection, the sensitivity of the *lytA*, *hylA*, and SP_2020 assays was 100%, whereas that of *piaB* was 93.3% (Table 1).

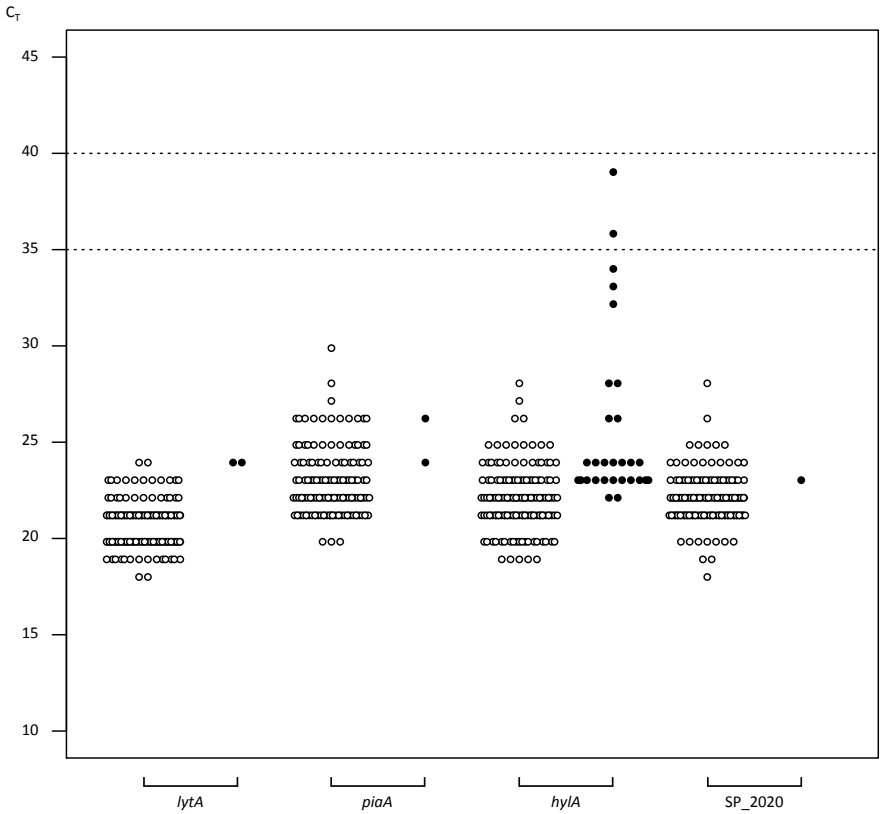


Figure 1. Real-time PCR detection of *S. pneumoniae* candidate target genes *lytA*, *piaB*, *hylA*, and SP_2020.

Open circles – *S. pneumoniae* control collection (n=150); closed circles – non-pneumococcal isolates: non-*S. pneumoniae spp.* control collection (n=31) and test collection (n=402). Number of samples amplifying for each gene: *lytA* – n=150 *S. pneumoniae* control collection and n=2 test collection; *piaB* – n=140 *S. pneumoniae* control collection and n=2 test collection; *hylA* – n=150 *S. pneumoniae* control collection, n=2 non-*S. pneumoniae spp*. control collection, and n=29 test collection; SP_2020 – n=150 *S. pneumoniae* control collection and n=1 test collection. Number of samples not amplifying each gene (not shown in the graphic – no C_T available): *lytA* – n=31 non-*S. pneumoniae spp.* control collection and n=400 test collection; *piaB* – n=10 *S. pneumoniae* control collection, n=31 non-*S. pneumoniae spp*. control collection, and n=400 test collection; *hylA* – n=29 non-*S. pneumoniae spp*. control collection and n=373 test collection; SP_2020 – n= 31 non-*S. pneumoniae spp.* control collection and n=401 test collection.

Table 1. Sensitivity, positive predictive value, and species misidentified by the real-time PCR assays tested.

PPV – positive predictive value

Assay (C_T≤35)	Sensitivity	PPV	Misidentified species (no. isolates out of 583)
lytA	100%	98.7%	*S. pseudopneumoniae* (2)
piaB	93.3%	98.6%	*S. pneumoniae* (10), *S. pseudopneumoniae* (2)
hylA	100%	83.8%	*S. mitis* (27), *S. oralis* (2)
SP_2020	100%	99.3%	Uncertain (1)
lytA+*piaB*	93.3%	100%	*S. pneumoniae* (10)
lytA+*hylA*	100%	100%	
lytA+SP_2020	100%	100%	
piaB+*hylA*	93.3%	100%	*S. pneumoniae* (10)
piBa+SP_2020	93.3%	100%	*S. pneumoniae* (10)
hylA+SP_2020	100%	100%	
lytA+*piaB*+*hylA*	93.3%	100%	*S. pneumoniae* (10)
lytA+*piaB*+SP_2020	93.3%	100%	*S. pneumoniae* (10)
lytA+*hylA*+SP_2020	100%	100%	
piaB+*hylA*+SP_2020	100%	100%	*S. pneumoniae* (10)
lytA+*piaB*+*hylA*+SP_2020	93.3%	100%	*S. pneumoniae* (10)

The absence of *piaB* has been previously reported for non-typeable pneumococci (Whalan *et al.*, 2006; Tavares *et al.*, 2014), but not for encapsulated pneumococci. Moreover, three of the encapsulated pneumococcal strains testing negative for *piaB* belong to serotypes targeted by pneumococcal conjugate vaccines. Therefore, according to our results, the identification of pneumococcus by detection of *piaB* could have consequences on studies aiming to evaluate vaccine efficacy.

To evaluate the positive predictive value (PPV) of the real-time PCR assays, 31 type strains of 23 *Streptococcus* species and 402 α-haemolytic non-pneumococcal isolates were tested. In total, 34 non-pneumococci (7.82%) were positive for at least one of the assays (Figure 1 and Table 1). In addition, two other non-pneumococci had a C_T value between 36 and 44. For these 36 isolates, DNA extraction was repeated and the assays were performed again for confirmation. In addition, MLSA was performed. All real-time PCR results were confirmed for the 36 isolates and all were assigned to non-pneumococcal *Streptococcus* species by MLSA (Figure 2 and Figure

3). The PPVs were, therefore, as follows: 98.7% for *lytA*, 98.6% for *piaB*, 83.8% for *hylA*, and 99.3% for SP_2020 (Table 1).

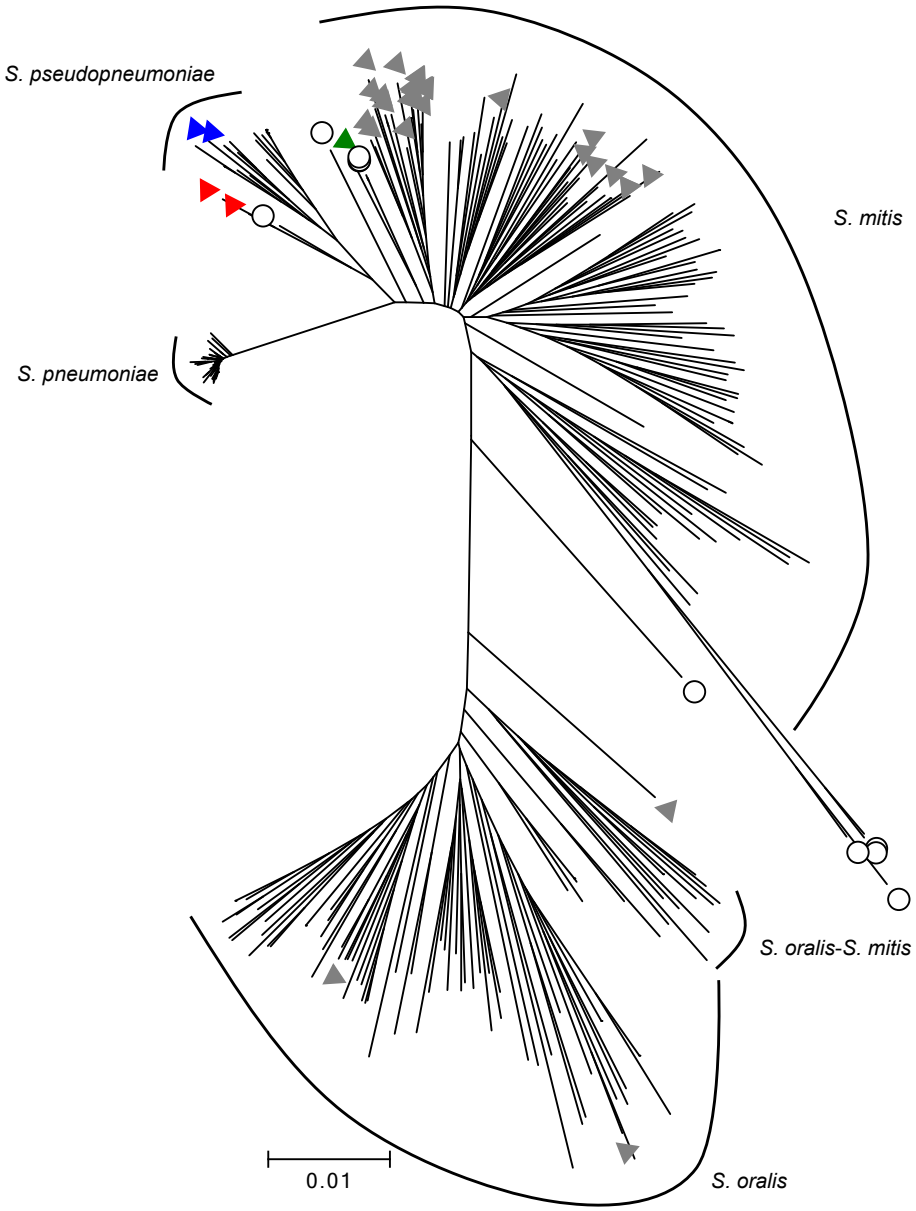


Figure 2. Phylogenetic tree based on concatenated MLSA sequences of the strains analysed in this study and *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* strains deposited at the eMLSA database.

A total of 34 non-pneumococcal isolates giving a positive result for at least one of the real-time PCR assays tested (plus two non-pneumococcal isolates with a C_T value between 36 and 44) were analysed. Triangles – strains analysed in this study: red – *lytA*$^+$; blue – *piaB*$^+$; grey – *hylA*$^+$; and green - SP_2020$^+$; open circles – strains deposited at the eMLSA database not assigned to the species level; species – as defined in the eMLSA database.

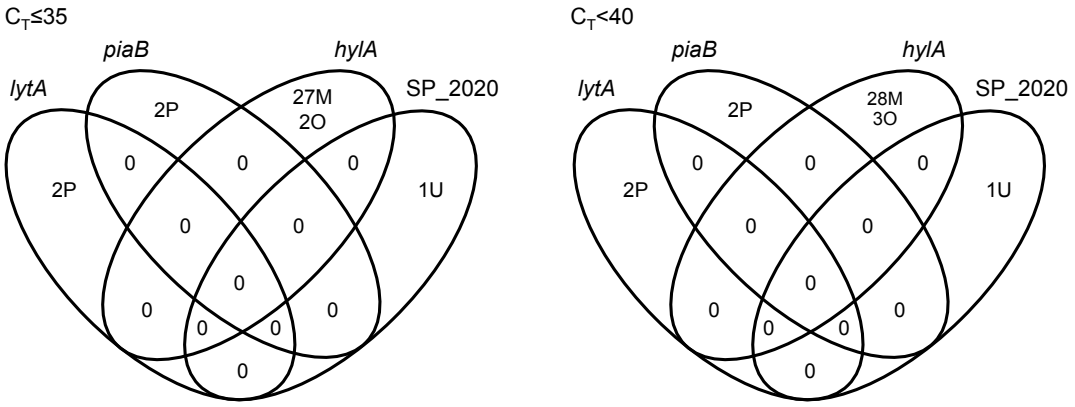


Figure 3. Number of non-pneumococcal positive samples for the real-time PCR assays tested.
A total of 31 type strains of 23 *Streptococcus* species and 402 α-haemolytic non-pneumococcal isolates were tested by the real-time PCR assays targeting *S. pneumoniae* genes *lytA*, *piaB*, *hylA*, and SP_2020. In this study, a sample was considered positive when the cycle threshold (C_T) value was equal or below 35 (left image). However, the usefulness of increasing the cut-off value to Ct<40, as proposed by others when using a combination of two real-time PCR assays, was also investigated (right image) (Trzcinski *et al.*, 2013; Wyllie *et al.*, 2014). P – *S. pseudopneumoniae*; M – *S. mitis*; O – *S. oralis*; U – uncertain.

The two isolates misidentified as pneumococcus by *lytA*-CDC were *S. pseudopneumoniae*. (Table 1) (Simões *et al.*, 2016). Homologues of *lytA* have been previously described in *S. pseudopneumoniae* and *S. mitis* (Whatmore *et al.*, 2000; Romero *et al.*, 2004; Shahinas *et al.*, 2011).

The assays with the best performance for the identification of pneumococcus were SP_2020 and *lytA*-CDC with a sensitivity of 100% and a PPV of 99.3% and 98.7%, respectively (p=0.564, Table 1). Analysis of all possible combinations between assays showed that it is possible to increase the PPV from 99.3% to 100% without decreasing the sensitivity by combining the SP_2020 assay with either the *lytA*-CDC or *hylA* assay. However, this increase was not statistically significant (p=0.315 for both).

In this study, an isolate was considered positive for a given assay when the C_T value was equal or below 35. This is the C_T value recommended by the WHO for the diagnosis of meningitis caused by pneumococcus using the *lytA*-CDC assay (WHO, 2011). Also according to the WHO procedure, a sample is considered negative if amplification occurs at C_T>40 and equivocal at C_T 36-40, needing to be retested after dilution in the latter case. On the other hand, for the detection of pneumococcus in carriage, a C_T<40 for both *lytA*-CDC and *piaB* assays ran in parallel has also been used to consider a sample positive for pneumococcus (Wyllie *et al.*, 2014). By combining two assays, it seems most likely that amplification at higher C_T would result from a lower concentration of pneumococcus in the sample rather than unspecific amplification of sequences of closely related species of the viridans group of *Streptococcus*. However, as only pure cultures of fixed concentration were used in this study, we opted for the more conservative approach. In such case, as only two samples had a C_T 36-40 (*hylA* assay), increasing the cut-off from C_T≤35 to C_T>40 was not found to increase the performance of the assays (Figure 3).

This study has a limitation. The test collection is biased towards non-pneumococcus with pneumococcal-like characteristics, as the isolates were initially selected as presumptive pneumococcus, and so does not represent the plethora of non-*S. pneumoniae Streptococcus spp.* present in the human microbiota. Although the determined PPV of the assays tested may be underestimated, this study clearly emphasizes the risk of using a single gene target for the identification of pneumococcus and identifies a novel promising pneumococcal target.

In conclusion, this study suggests that detection of SP_2020 in combination with *lytA*-CDC is a powerful strategy for the identification of pneumococcus in polymicrobial samples.

Acknowledgements

This work was funded by Project LISBOA-01-0145-FEDER (Microbiologia Molecular, Estrutural e Celular) funded by FEDER funds through COMPETE2020 – Programa Operacional Competitividade e Internacionalização (POCI) and by national funds through FCT – Fundação para a Ciência e a Tecnologia, Portugal (grants PTDC/BIA-MIC/64010/2006 and PTDC/BIA-BEC/098289/2008 to RSL, SFRH/BD/70147/2010 to DAT, and UID/CBO/04612/2013).

The authors are thankful to Ana Cristina Paulo from the Molecular Microbiology of Human Pathogens Laboratory, ITQB-NOVA for her support with the statistical analysis and preparation of Figure 1. The authors thank Dea Shahinas from the University of Toronto, Canada for kindly providing *S. pseudopneumoniae* strain IS7943.

References

Arbique, J.C., C. Poyart, P. Trieu-Cuot, G. Quesne, G. Carvalho Mda, A.G. Steigerwalt, R.E. Morey, D. Jackson, R.J. Davidson and R.R. Facklam. (2004). Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. *J Clin Microbiol* **42**, 4686-96.

Bishop, C.J., D.M. Aanensen, G.E. Jordan, M. Kilian, W.P. Hanage and B.G. Spratt. (2009). Assigning strains to bacterial species via the internet. *BMC Biol* **7**, 3.

Brown, J.S., S.M. Gilliland and D.W. Holden. (2001). A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol* **40**, 572-85.

Carvalho Mda, G., M.L. Tondella, K. McCaustland, L. Weidlich, L. McGee, L.W. Mayer, A. Steigerwalt, M. Whaley, R.R. Facklam, B. Fields, G. Carlone, E.W. Ades, R. Dagan and J.S. Sampson. (2007). Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *J Clin Microbiol* **45**, 2460-6.

del Amo, E., L. Selva, M.F. de Sevilla, P. Ciruela, P. Brotons, M. Trivino, S. Hernandez, J.J. Garcia-Garcia, A. Dominguez and C. Munoz-Almagro. (2015). Estimation of the invasive disease potential of *Streptococcus pneumoniae* in children by the use of direct capsular typing in clinical specimens. *Eur J Clin Microbiol Infect Dis* **34**, 705-11.

Feldman, C., R. Cockeran, M.J. Jedrzejas, T.J. Mitchell and R. Anderson. (2007). Hyaluronidase augments pneumolysin-mediated injury to human ciliated epithelium. *Int J Infect Dis* **11**, 11-5.

Leisenring, W., T. Alonzo and M.S. Pepe. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* **56**, 345-51.

Llull, D., R. Lopez and E. Garcia. (2006). Characteristic signatures of the *lytA* gene provide a basis for rapid and reliable diagnosis of *Streptococcus pneumoniae* infections. *J Clin Microbiol* **44**, 1250-6.

Mundy, L.S., E.N. Janoff, K.E. Schwebke, C.J. Shanholtzer and K.E. Willard. (1998). Ambiguity in the identification of *Streptococcus pneumoniae*. Optochin, bile solubility, quellung, and the AccuProbe DNA probe tests. *Am J Clin Pathol* **109**, 55-61.

Obert, C., J. Sublett, D. Kaushal, E. Hinojosa, T. Barton, E.I. Tuomanen and C.J. Orihuela. (2006). Identification of a Candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun* **74**, 4766-77.

Pozzi, G., M.R. Oggioni and A. Tomasz. (1989). DNA probe for identification of *Streptococcus pneumoniae*. *J Clin Microbiol* **27**, 370-2.

Richter, S.S., K.P. Heilmann, C.L. Dohrn, F. Riahi, S.E. Beekmann and G.V. Doern. (2008). Accuracy of phenotypic methods for identification of *Streptococcus pneumoniae* isolates included in surveillance programs. *J Clin Microbiol* **46**, 2184-8.

Romero, P., R. Lopez and E. Garcia. (2004). Characterization of LytA-like N-acetylmuramoyl-L-alanine amidases from two new *Streptococcus mitis* bacteriophages provides insights into the properties of the major pneumococcal autolysin. *J Bacteriol* **186**, 8229-39.

Sá-Leão, R., S. Nunes, A. Brito-Avô, N. Frazão, A.S. Simões, M.I. Crisóstomo, A.C. Paulo, J. Saldanha, I. Santos-Sanches and H. de Lencastre. (2009). Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. *Clin Microbiol Infect* **15**, 1002-7.

Satzke, C., P. Turner, A. Virolainen-Julkunen, P.V. Adrian, M. Antonio, K.M. Hare, A.M. Henao-Restrepo, A.J. Leach, K.P. Klugman, B.D. Porter, R. Sa-Leao, J.A. Scott, H. Nohynek, K.L. O'Brien and W.H.O.P.C.W. Group. (2013). Standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*: updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. *Vaccine* **32**, 165-79.

Shahinas, D., G.S. Tamber, G. Arya, A. Wong, R. Lau, F. Jamieson, J.H. Ma, D.C. Alexander, D.E. Low and D.R. Pillai. (2011). Whole-genome sequence of *Streptococcus pseudopneumoniae* isolate IS7493. *J Bacteriol* **193**, 6102-3.

Simões, A.S., R. Sá-Leão, M.J. Eleveld, D.A. Tavares, J.A. Carriço, H.J. Bootsma and P.W. Hermans. (2010). Highly penicillin-resistant multidrug-resistant pneumococcus-like strains colonizing children in Oeiras, Portugal: genomic characteristics and implications for surveillance. *J Clin Microbiol* **48**, 238-46.

Simões, A.S., D.A. Tavares, D. Rolo, C. Ardanuy, H. Goossens, B. Henriques-Normark, J. Linares, H. de Lencastre and R. Sá-Leão. (2016). *lytA*-based identification methods can misidentify *Streptococcus pneumoniae*. *Diagn Microbiol Infect Dis*

Simões, A.S., C. Valente, H. de Lencastre and R. Sá-Leão. (2011). Rapid identification of noncapsulated *Streptococcus pneumoniae* in nasopharyngeal samples allowing detection of co-colonization and reevaluation of prevalence. *Diagn Microbiol Infect Dis* **71**, 208-16.

Tavares, D.A., A.S. Simões, H.J. Bootsma, P.W. Hermans, H. de Lencastre and R. Sá-Leão. (2014). Non-typeable pneumococci circulating in Portugal are of *cps* type NCC2 and have genomic features typical of encapsulated isolates. *BMC Genomics* **15**, 863.

Trzcinski, K., D. Bogaert, A. Wyllie, M.L. Chu, A. van der Ende, J.P. Bruin, G. van den Dobbelsteen, R.H. Veenhoven and E.A. Sanders. (2013). Superiority of trans-oral over trans-nasal sampling in detecting *Streptococcus pneumoniae* colonization in adults. *PLoS One* **8**, e60520.

Wester, C.W., D. Ariga, C. Nathan, T.W. Rice, J. Pulvirenti, R. Patel, F. Kocka, J. Ortiz and R.A. Weinstein. (2002). Possible overestimation of penicillin resistant *Streptococcus pneumoniae* colonization rates due to misidentification of oropharyngeal streptococci. *Diagn Microbiol Infect Dis* **42**, 263-8.

Whalan, R.H., S.G. Funnell, L.D. Bowler, M.J. Hudson, A. Robinson and C.G. Dowson. (2006). Distribution and genetic diversity of the ABC transporter lipoproteins PiuA and 3A within *Streptococcus pneumoniae* and related streptococci. *J Bacteriol* **188**, 1031-8.

Whatmore, A.M., A. Efstratiou, A.P. Pickerill, K. Broughton, G. Woodard, D. Sturgeon, R. George and C.G. Dowson. (2000). Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "Atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. *Infect Immun* **68**, 1374-82.

WHO, 2011. PCR for detection and characterization of bacterial meningitis pathogens: *Neisseria meningitidis*, *Haemophilus influenzae*, and *Streptococcus pneumoniae*. In: Laboratory Methods for the Diagnosis of Meningitis caused by *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Haemophilus influenzae*, WHO Manual, 2nd edition.

Wyllie, A.L., M.L. Chu, M.H. Schellens, J. van Engelsdorp Gastelaars, M.D. Jansen, A. van der Ende, D. Bogaert, E.A. Sanders and K. Trzcinski. (2014). *Streptococcus pneumoniae* in saliva of Dutch primary school children. *PLoS One* **9**, e102045.

Chapter 5

Concluding remarks

The work developed under the scope of this thesis enabled us to gain further insights into the biology of non-typeable pneumococci (NT), a group of pneumococcal strains that has for long been largely ignored. Moreover, the usefulness of *lytA*-based pneumococcal identification methods was assessed and new candidate targets were tested for the identification of pneumococcus by real-time PCR.

The results presented in **Chapter 2** provide strong evidence that non-typeable pneumococci are very similar to encapsulated pneumococci in terms of core genome. The major differences found, apart from the lack of capsule, was the absence of type-I and type-II pili, choline-binding protein A, and pneumococcal surface protein A from all NT tested. Furthermore, NT carriage strains circulating in Portugal over a decade were found to be a homogeneous group belonging to *cps* type NCC2, i.e., harbouring *aliB*-like ORF1 and *aliB*-like ORF2. Given the fact that NT are not included in currently available vaccines and its prevalence in colonisation is increasing, we propose that NT should be routinely identified and reported in pneumococcal carriage surveillance studies.

In **Chapter 3** we characterised 11 α-haemolytic streptococcal isolates of uncertain species identification by methods widely used to identify pneumococcus such as optochin susceptibility, bile solubility, MLST and *lytA*-BsaAI-RFLP. Of particular interest, a pneumococcal isolate had a *lytA* homologue and lacked the typical pneumococcal *lytA* and two *S. pseudopneumoniae* isolates had the characteristic pneumococcal *lytA*. These three isolates were misidentified by two methods commonly used to distinguish between pneumococcus and closely related species: the *lytA*-BsaAI-RFLP and the *lytA*-CDC real-time PCR assay. These results should raise

awareness for the existence of both pneumococci and non-pneumococci that can be misidentified by currently widely accepted *lytA*-based identification methods.

Following the results presented in Chapter 3, in **Chapter 4** we evaluated the performance of four real-time PCR assays for the identification of pneumococcus. Using a collection of close to 600 α-haemolytic streptococcal isolates, including 150 pneumococci and 31 strains from other *Streptococcu*s species, we evaluated: i) the *lytA*-CDC real-time PCR assay; ii) a *piaB* real-time PCR assay that hass been used in parallel with *lytA* to increase specificity; and iii) two new candidate targets, *hylA* and SP_2020. The real-time PCR assay with the best performance in identifying pneumococcus were the SP_2020 and *lytA*-CDC assays with a positive predictive value of 99.3% and 98.7%, respectively (p=0.564).

Altogether, the work presented in this thesis contributed to improve knowledge about NT and pneumococcal molecular-based identification methods. Yet, some interesting topics related with these studies were left untouched and could be the focus of future investigation.

First, to study the genomic content of NT we have used comparative genomic hybridisation (CGH) which is a valid method but that bears some intrinsic limitations. Using CGH, strains can only be characterised in comparison to genes represented in the microarray. Although we used a microarray representative of 10 pneumococcal strains, which should have overcome this limitation to some extent, our conclusions may have been limited. We may have missed, for example, parts of the accessory genome of NT. Also, some pneumococcal virulence genes that were found absent from the NT strains studied, such as choline binding protein A and pneumococcal

surface protein A, could, in fact, be present in a divergent form (Valentino *et al.*, 2014). An alternative could be the use of WGS. In fact, a recent study of a global collection of NT found bigger accessory genomes and genome sizes in 'classical' NT strains when compared with NT strains related to encapsulated strains (Hilty *et al.*, 2014).

One interesting question regarding the biology of NT is whether these strains would have the potential to cause (invasive) disease if shielded by a pneumococcal polysaccharide capsule. We have shown, under the scope of this thesis, that NT harbour most pneumococcal virulence genes. It is also known that NT are usually resistant to antimicrobials and, although rarely, capable of causing invasive disease (Sá-Leão *et al.*, 2006; Hilty *et al.*, 2014; Park *et al.*, 2014). Because NT are highly transformable and capsular switching can occur between pneumococci (Weiser and Kapoor, 1999; Wyres *et al.*, 2013), it is reasonable to speculate that NT could potentially acquire a pneumococcal capsule. It would be interesting to estimate the frequency at which NT are able to acquire a pneumococcal capsule *in vitro* and if this capsule could be maintained or whether the genotype would soon be reverted. It would also be interesting to estimate the invasive disease potential of strains with NT backgrounds but expressing a capsule.

Regarding the identification of NT, we have previously described a multiplex PCR scheme targeting 16S rRNA, *cpsA*, *lytA*, and *aliB*-like ORF2 genes to distinguish NT from closely related *Streptococcus* species (Simões *et al.*, 2011). Although very useful to detect NT strains of *cps* type NCC2, this method has not been designed to detect strains of the recently described *cps* type NCC1 (Park *et al.*, 2012; Salter *et al.*, 2012).

This group of NT have been shown to harbour the *pspK*/*nspA* gene at the capsular region. Thus, it would be relevant to update this multiplex PCR scheme (by including primers targeting *pspK*) to enable direct detection of NT strains of *cps* type NCC1.

In Chapter 3, a pneumococcal isolate harbouring a homologue of *lytA* and two *S. pseudopneumoniae* harbouring *lytA* were described. It would be very interesting to characterise these isolates by WGS to investigate whether only the *lytA* has been lost/gained or if a smooth transition between pneumococcus and closely related species can be observed in these strains as it has been previously proposed to occur for some isolates (Hakenbeck *et al.*, 2001). To address this question, it would also be interesting to systematically screen for isolates of ambiguous identification and to perform a WGS study aiming to characterise a large collection of such isolates. This could potentially increase the knowledge about the evolution of pneumococcus and closely related species generated by the analysis of pneumococcal, *S. mitis*, *S. pseudopneumoniae*, *S. oralis*, and *S. infantis* genomes (Kilian *et al.*, 2014).

Finally, but not less importantly, the SP_2020 real-time PCR assay for the identification of pneumococcus developed under the scope of this thesis was only tested for pure cultures. It would be useful to extend the tests to complex carriage samples from both children and adults, and to disease samples from several anatomical sites. This could be done after automated DNA extraction directly from both carriage and clinical samples or after culture enrichment for pneumococcus (as the latter approach has been shown to increase the sensitivity of molecular detection of pneumococcus) (da Gloria Carvalho *et al.*, 2010; Wyllie *et al.*, 2014).

In children, carriage loads are higher and false positives by a real-time PCR targeting the pneumococcal pneumolysin have been obtained when testing serum samples of healthy children (Dagan *et al.*, 1998). Although similar studies using the *lytA*-CDC real-time PCR assay did not reach the same results, as no false positive urine or blood samples were detected in healthy children, it is important to address this issue for any new proposed assay (Azzari *et al.*, 2011; Rouphael *et al.*, 2011). For these reasons, it would be interesting to perform a longitudinal study on children with invasive pneumococcal disease, where the identification of pneumococcus by the SP_2020 real-time PCR assay would be performed both in blood/cerebrospinal fluid and urine samples at multiple time points, for example: when the child has a positive blood/cerebrospinal fluid culture for pneumococcus, during the recovery period, at hospital discharge, and after hospital discharge. Data from invasive samples should then be compared with data from carriage samples from matched controls. Such a study would enable the estimation of false positives arising from the presence of pneumococcal DNA in normally sterile sites outside the timeframe of disease. Of course similar studies could also be performed for pneumonia and otitis media.

Overall, the studies presented in this thesis represent an important contribution to the knowledge of (i) NT, a poorly studied group of pneumococcus that may be crucial to the species as the reservoir of the pneumococcal gene pool and the driving force of pneumococcal evolution (Chewapreecha *et al.*, 2014; Hilty *et al.*, 2014) and (ii) the identification of pneumococcus by molecular methods, namely real-time PCR.

References

Azzari, C., M. Cortimiglia, M. Moriondo, C. Canessa, F. Lippi, F. Ghiori, L. Becciolini, M. de Martino and M. Resti. (2011). Pneumococcal DNA is not detectable in the blood of healthy carrier children by real-time PCR targeting the *lytA* gene. *J Med Microbiol* **60**, 710-4.

Chewapreecha, C., S.R. Harris, N.J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D.M. Aanensen, A.E. Mather, A.J. Page, S.J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner and S.D. Bentley. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46**, 305-9.

da Gloria Carvalho, M., F.C. Pimenta, D. Jackson, A. Roundtree, Y. Ahmad, E.V. Millar, K.L. O'Brien, C.G. Whitney, A.L. Cohen and B.W. Beall. (2010). Revisiting pneumococcal carriage by use of broth enrichment and PCR techniques for enhanced detection of carriage and serotypes. *J Clin Microbiol* **48**, 1611-8.

Dagan, R., O. Shriker, I. Hazan, E. Leibovitz, D. Greenberg, F. Schlaeffer and R. Levy. (1998). Prospective study to determine clinical relevance of detection of pneumococcal DNA in sera of children by PCR. *J Clin Microbiol* **36**, 669-73.

Hakenbeck, R., N. Balmelle, B. Weber, C. Gardes, W. Keck and A. de Saizieu. (2001). Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. *Infect Immun* **69**, 2477-86.

Hilty, M., D. Wuthrich, S.J. Salter, H. Engel, S. Campbell, R. Sá-Leão, H. de Lencastre, P. Hermans, E. Sadowy, P. Turner, C. Chewapreecha, M. Diggle, G. Pluschke, L. McGee, O. Koseoglu Eser, D.E. Low, H. Smith-Vaughan, A. Endimiani, M. Kuffer, M. Dupasquier, E. Beaudoing, J. Weber, R. Bruggmann, W.P. Hanage, J. Parkhill, L.J. Hathaway, K. Muhlemann and S.D. Bentley. (2014). Global phylogenomic analysis of nonencapsulated *Streptococcus pneumoniae* reveals a deep-branching classic lineage that is distinct from multiple sporadic lineages. *Genome Biol Evol* **6**, 3281-94.

Kilian, M., D.R. Riley, A. Jensen, H. Bruggemann and H. Tettelin. (2014). Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. *MBio* **5**, e01490-14.

Park, I.H., K.A. Geno, L.K. Sherwood, M.H. Nahm and B. Beall. (2014). Population-based analysis of invasive nontypeable pneumococci reveals that most have defective capsule synthesis genes. *PLoS One* **9**, e97825.

Park, I.H., K.H. Kim, A.L. Andrade, D.E. Briles, L.S. McDaniel and M.H. Nahm. (2012). Nontypeable pneumococci can be divided into multiple *cps* types, including one type expressing the novel gene *pspK*. *MBio* **3**,

Rouphael, N., S. Steyn, M. Bangert, J.S. Sampson, P. Adrian, S.A. Madhi, K.P. Klugman and E.W. Ades. (2011). Use of 2 pneumococcal common protein real-time polymerase chain reaction assays in healthy children colonized with *Streptococcus pneumoniae*. *Diagn Microbiol Infect Dis* **70**, 452-4.

Sá-Leão, R., A.S. Simões, S. Nunes, N.G. Sousa, N. Frazão and H. de Lencastre. (2006). Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. *Microbiology* **152**, 367-76.

Salter, S.J., J. Hinds, K.A. Gould, L. Lambertsen, W.P. Hanage, M. Antonio, P. Turner, P.W. Hermans, H.J. Bootsma, K.L. O'Brien and S.D. Bentley. (2012). Variation at the capsule locus, *cps*, of mistyped and non-typable *Streptococcus pneumoniae* isolates. *Microbiology* **158**, 1560-9.

Simões, A.S., C. Valente, H. de Lencastre and R. Sá-Leão. (2011). Rapid identification of noncapsulated *Streptococcus pneumoniae* in nasopharyngeal samples allowing detection of co-colonization and reevaluation of prevalence. *Diagn Microbiol Infect Dis* **71**, 208-16.

Valentino, M.D., A.M. McGuire, J.W. Rosch, P.J. Bispo, C. Burnham, C.M. Sanfilippo, R.A. Carter, M.E. Zegans, B. Beall, A.M. Earl, E.I. Tuomanen, T.W. Morris, W. Haas and M.S. Gilmore. (2014). Unencapsulated *Streptococcus pneumoniae* from conjunctivitis encode variant traits and belong to a distinct phylogenetic cluster. *Nat Commun* **5**, 5411.

Weiser, J.N. and M. Kapoor. (1999). Effect of intrastrain variation in the amount of capsular polysaccharide on genetic transformation of *Streptococcus pneumoniae*: implications for virulence studies of encapsulated strains. *Infect Immun* **67**, 3690-2.

Wyllie, A.L., M.L. Chu, M.H. Schellens, J. van Engelsdorp Gastelaars, M.D. Jansen, A. van der Ende, D. Bogaert, E.A. Sanders and K. Trzcinski. (2014). *Streptococcus pneumoniae* in saliva of Dutch primary school children. *PLoS One* **9**, e102045.

Wyres, K.L., L.M. Lambertsen, N.J. Croucher, L. McGee, A. von Gottberg, J. Linares, M.R. Jacobs, K.G. Kristinsson, B.W. Beall, K.P. Klugman, J. Parkhill, R. Hakenbeck, S.D. Bentley and A.B. Brueggemann. (2013). Pneumococcal capsular switching: a historical perspective. *J Infect Dis* **207**, 439-49.

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100 | Fax (+351) 214 411 277

www.itqb.unl.pt